La validez en los exámenes de alto impacto
Un enfoque desde la lógica argumentativa

Arturo Mendoza Ramos*

La validez en la evaluación educativa de alto impacto representa un desafío para las instituciones que se dedican a diseñar y validar exámenes; sin embargo, los organismos internacionales que han provisto estándares de calidad para la evaluación no proporcionan en realidad una guía clara de cómo elaborar exámenes, ni la manera mediante la cual se pueda proveer una evaluación justa y equitativa para los sustentantes. En el presente artículo se expone la importancia de la validez a partir de la segunda mitad del siglo XX, las dificultades latentes al momento de validar exámenes a finales del siglo pasado y, posteriormente, se presenta una novedosa concepción de abordar la validez en la evaluación educativa. Vista desde la lógica de la argumentación, la validez consiste en todo un proceso de inferencias que deben encontrarse sustentadas en la evidencia y en el respaldo de las mismas. Este modelo para validar exámenes desde la lógica argumentativa se está siguiendo en diversas áreas educativas a nivel internacional; sin embargo, ha cobrado especial relevancia en la evaluación de segundas lenguas.

Palabras clave

Evaluación educativa
Evaluación de modelos
Evaluación de pruebas
Validez de las pruebas
Exámenes

Validity of high impact education assessment represents a challenge for institutions that design and validate exams. However, the international bodies that have provided quality standards in assessment in fact do not provide clear guidelines regarding how to develop exams, nor how to provide exam candidates with fair and equitable evaluations. This paper presents the importance validity has adopted since the second half of the twentieth century, the latent difficulties faced when validating exams at the end of the last century, as well as an innovative idea to approach the validity of education assessment. Seen through the lens of argumentative logic, validity consists of a whole process of inferences that must be evidence-based. This exam validation model based on argumentative logic is being applied in various educational areas internationally, and has become particularly germane to second language evaluation.

Keywords

Education assessment
Assessment models
Test assessment
Test validity
Tests

Recepción: 9 de junio de 2014 | Aceptación: 29 de agosto de 2014

* Profesor asociado "C" tiempo completo en el Centro de Enseñanza de Lenguas Extranjeras de la Universidad Nacional Autónoma de México (UNAM). Maestro en Lingüística Aplicada y estudiante del Doctorado en Lingüística de la Universidad Nacional Autónoma de México (UNAM). Línea de investigación: evaluación y certificación de lenguas. Publicaciones recientes: (2015), "La selección de las tareas de escritura en los exámenes de lengua extranjera destinados al ámbito académico", *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, vol. 9, núm. 11, en: http://www.nebrija.com/revista-linguistica/files/articulosPDF/articulo_5500062b360e6.pdf; (2014), "Las prácticas de evaluación docente y las habilidades de escritura requeridas en el nivel posgrado", *Innovación Educativa*, vol. 14, núm. 66, pp. 147-175. CE: a.mendoza@cele.unam.mx

Introducción

> If validity is to be considered the most fundamental consideration in developing and evaluating tests, it needs to address the suitability of the test for its intended function (Kane, 2013: 62).

En el ámbito educativo en México, la evaluación de alto impacto[1]que llevan a cabo organizaciones nacionales e internacionales ha aumentado considerablemente; esto se debe a la necesidad de contar con mecanismos estandarizados que permitan servir como puntos de comparación, tanto entre estudiantes connacionales como entre nacionales y extranjeros. En el primer caso, basta mencionar los exámenes de ingreso a la educación media superior (EXANI) o superior de las universidades públicas del área metropolitana, ya que cientos de miles de aspirantes presentan dichos exámenes cada año. En el caso de los exámenes internacionales destaca, por ejemplo, la prueba diseñada por el Programa Internacional de Evaluación de Estudiantes (PISA, por sus siglas en inglés), la cual tiene como objetivo medir las competencias que poseen los jóvenes de 15 años en matemáticas, ciencias y lectura en más de 60 países.

Debido a que los exámenes de alto impacto juegan un papel primordial en la vida de los examinados, es imprescindible que tales exámenes evalúen de manera justa y equitativa a los candidatos que cada año los realizan; es por ello que en la evaluación, la ética ha adquirido un lugar tan importante como la validez y confiabilidad de la evaluación misma. Una de las dificultades que ha girado en torno a la noción de validez se refiere a que los estándares internacionales (APA, AERA y NCME)[2] la conciben como una serie de elementos disgregados y desarticulados que resultan difíciles de incorporar al momento de diseñar y validar exámenes; por tal motivo, desde comienzos del siglo XXI la validez se ha concebido como un argumento que requiere de ciertas inferencias, evidencias y sustento que respalde la toma de decisiones.

En el presente texto se presenta, en primer lugar, una breve reseña de los orígenes de la noción de validez en los exámenes, partiendo de los conceptos aportados por Cronbach (1988) y Messick (1989); también se aborda el esfuerzo por incluir estos conceptos dentro de los modelos de diseño y de validación en los exámenes durante la década de los años noventa. Más adelante se describe la problemática que se suscitó a finales del siglo XX al tratar de poner en práctica tanto los modelos teóricos para validar exámenes, como los conceptos de validez fragmentados que eran utilizados por los expertos al evaluar exámenes de alto impacto. Posteriormente, se describe el modelo de argumentación de Toulmin (1958) y la aplicación que ha tenido en la evaluación para la generación de diversos enfoques de validación de exámenes. Más adelante, se discuten las ventajas y desventajas de cada uno de los enfoques y se presenta la aplicación de los conceptos teóricos en un esquema diseñado para validar los exámenes de comprensión de lectura del Centro de Enseñanza de Lenguas Extranjeras de la UNAM. Por último, se ofrece una breve conclusión de la validez, vista desde la óptica de la lógica argumentativa.

Los orígenes de la validez en los exámenes

La validez de cualquier examen de alto impacto es indispensable para tomar decisiones justas y equitativas con base en información

1 Un examen de alto impacto es aquel cuyo resultado es empleado para tomar decisiones importantes que conllevan consecuencias positivas o negativas en quienes los sustentan. A diferencia de los exámenes a gran escala, los de alto impacto pueden o no ser administrados a grandes poblaciones. Por ejemplo, en el contexto educativo, los resultados de los exámenes de aptitud, certificación y dominio son frecuentemente empleados por las partes interesadas como criterios de egreso/ingreso tanto en la educación media superior como superior.

2 APA: American Psychological Association; AERA: American Educational Research Association; NCME: National Council on Measurement in Education.

veraz respecto de las habilidades/competencias que posee un sustentante. La validez se define antes, durante y después de la administración de un examen, y su fin, en términos generales, es indagar si un examen en realidad mide lo que debería medir; sin embargo, la manera de definir la validez ha variado sustancialmente desde su florecimiento —a mediados del siglo XX— hasta nuestros días.

Durante la primera mitad del siglo XX, el concepto de validez se sustentaba en modelos estadísticos basados en el criterio de los exámenes; es decir, la validez era definida como la correlación entre el puntaje de un examen y el puntaje definido como criterio. Este criterio, por lo general, se encontraba dado por exámenes similares que evaluaban las mismas habilidades; por lo tanto, un examen se consideraba como válido exclusivamente si medía lo mismo que otros instrumentos ya existentes.

Cronbach y Meehl introdujeron cuatro tipos de validez: predictiva, concurrente, de contenido y de constructo, los cuales constituyeron los ejes fundamentales de la validez de los exámenes hasta finales del siglo XX. Dentro de estas categorías, destaca la validez de constructo, la cual, según los autores, "se ve involucrada cuando un examen ha de ser interpretado como una medida de algún atributo o cualidad que no es definida operacionalmente" (1955: 282). Esta idea de constructo partió de la necesidad de concebir la validez en términos científicos, es decir, en la definición del constructo desde la teoría y, por lo tanto, en la convicción de que la interpretación del constructo puede ser evaluada empíricamente junto con la teoría que lo sustenta.

En 1974, los organismos estadounidenses encargados de desarrollar los estándares a considerar en la elaboración de exámenes (APA, AERA y NCME), definieron los cuatro tipos de validez descritos por Cronbach y Meehl (1955): predictiva, concurrente, de contenido y de constructo. La problemática que se suscitó a finales de esta década fue que los diferentes métodos de validez eran tratados como diversos instrumentos que podían ser empleados en distintas situaciones de acuerdo a como los evaluadores consideraban que fuera más conveniente emplearlas. El punto fundamental era que no se contaba con un marco general de validación para los exámenes, puesto que la validez de contenido y concurrente dependían del tipo de examen que se quería validar. Consecuentemente, debido a que la validez de constructo era la única que contaba con un sustento científico, se adoptó como marco para la validación de los exámenes.

Pese a que la validez de constructo propuesta por Cronbach y Meehl (1955) causaba dificultades en su aplicación, uno de los mayores logros de los autores fue demostrar que la interpretación de los resultados no podía darse por hecho, y que se necesitaba de una serie de afirmaciones circunscritas en un marco teórico. En realidad, una de las grandes aportaciones de Cronbach y Meehl al concepto de validez se debió a su concepción no como un mecanismo matemático, como es el caso de la confiabilidad, sino más bien como todo un proceso denominado "validez de argumento". Posteriormente, éste fue retomado por Kane (1992) en su enfoque de validación para los exámenes, que consiste en recabar evidencia para llevar a cabo inferencias, más que exclusivamente para validar instrumentos.

Gracias al gran trabajo de Cronbach a lo largo de varias décadas, se cimentaron las bases que propiciaron el surgimiento del modelo de validez de Samuel Messick en 1989. Este modelo, aún vigente, es el más representativo; sobre él versan las discusiones que hoy en día existen en materia de validez en la evaluación de exámenes de alto impacto. Messick definió este término de la siguiente manera: "*Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment*" (1989: 13).

Sin lugar a dudas, la mayor aportación de Messick fue concebir la validez como un aspecto unitario y no como una serie de elementos disgregados y desarticulados. Como se puede apreciar en el Cuadro 1, desde la óptica de Messick la evaluación es un procedimiento para extraer inferencias (toma de decisiones) a partir de ciertas evidencias (exámenes); de tal suerte que el objetivo primordial de la validez empírica de los exámenes será demostrar que las interpretaciones y la toma de decisiones respecto al desempeño de un sustentante en un examen de dominio son justas y equitativas.

Cuadro 1. Facetas de validez como una matriz progresiva

	Test interpretation	Test use
Evidential basis	Construct Validity (CV)	CV + Relevance / Utility (R/U)
Consequential basis	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

Fuente: tomado de Messick, 1995: 748.

En suma, una de las grandes aportaciones de Cronbach y Messick fue concebir la validez como un proceso de argumentación, lo cual tuvo gran influencia en los enfoques de validez de la primera década del siglo XXI. No obstante, aun cuando el modelo de Messick (1989) constituyó un parteaguas en la noción unitaria de validez al incluir las consecuencias y el uso del examen dentro de su matriz de facetas de validez, una de las mayores dificultades de este modelo se debió a su materialización; en otras palabras, a lo abstracto y a la poca claridad como guía para aplicarlo al proceso de validación de exámenes (Bachman, 2005; Kane, 1992; Mislevy *et al*., 2003). Por tal razón, a finales del siglo XX y comienzos del XXI, la directriz para validar exámenes consistía, más bien, en una serie de características necesarias dentro de la validez, pero que no constituían un procedimiento lógico mediante el cual se encontraban vinculadas y relacionadas unas con otras (Bachman, 2005).

El panorama de la validez a finales del siglo XX y comienzos del XXI

Pese a las dificultades para materializar el modelo de Messick, la inclusión del significado social, los valores, la ética y las consecuencias de la evaluación fomentaron la necesidad de incorporarlos dentro del diseño y validación de exámenes. Por estos motivos, a finales del siglo XX, debido a la falta de integración e interrelación entre los diversos conceptos de validez, el derrotero de la validez dentro del ámbito de la evaluación comenzó a orientarse hacia la incorporación de un argumento que fungiera como "paraguas", y que no sólo describiera los componentes de validez, sino que además los integrara e interrelacionara entre sí. De ahí que el modelo de argumentación de Toulmin (1958) fuera retomado por los teóricos que a lo largo de dos décadas habían incorporado dicho modelo con el fin de concebir la validez de forma más global que fragmentaria. Por lo anterior, antes de describir los enfoques más influyentes en la validez de exámenes de alto impacto, será necesario describir el modelo de argumentación de Toulmin.

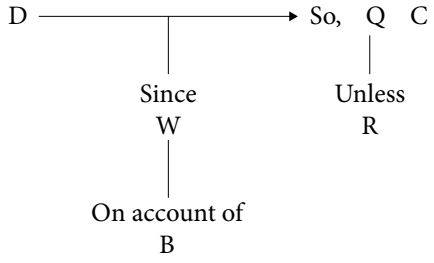
El modelo de argumentación de Toulmin (1958) y Toulmin, Janik y Rieke (1979; 1984)

El modelo de argumentación propuesto por Toulmin (1958), revisado y ampliado por Toulmin *et al.* (1979; 1984), ha tenido aplicaciones en diversos ámbitos como el derecho,

las ciencias duras, la medicina y la estética; no obstante, no fue sino hasta la década de los noventa del siglo pasado cuando comenzó a contemplarse su adecuación dentro del proceso de validación de exámenes (Kane, 1992).

Toulmin (1958) considera como punto de partida en el proceso de argumentación dos componentes: una aserción (*claim*) y los hechos o datos (*data*) que fungen como base o fundamento para validar dicha aserción (Fig. 1). Sin embargo, el verdadero reto, afirma el autor, no se encuentra representado por la aserción que se establece, sino por el procedimiento que se lleva a cabo para llegar a ella. Partiendo de esta premisa, el autor menciona que lo importante no es fortalecer nuestros datos, sino las proposiciones que de ellos podamos derivar: reglas, principios o inferencias. A este tipo de inferencias las denomina garantías (*warrants*); las cuales, aunque puedan parecen obviedades, en realidad no lo son.

Figura 1. Modelo de argumentación de Toulmin



Fuente: tomado de Toulmin, 1958: 104.

D - (Data) Datos; W - (Warrants) Garantías; B - (Backing) Respaldo; Q - (Qualifiers) Calificadores modales; R - (Rebuttal) Refutación; R - (Claim) Aserción.*

* Toulmin (1958: 105) proporciona el siguiente ejemplo con base en su modelo: Harry nació en Bermudas (datos); por lo tanto, es británico (aserción), debido a que los hombres que nacen en Bermudas son británicos (garantías) con base en los estatutos y disposiciones legales (fundamentos). Lo anterior sería contundentemente cierto, siempre y cuando (refutación) los padres de Harry no fueran extranjeros o él se hubiera naturalizado estadounidense, por ejemplo. El autor proporciona este ejemplo para distinguir los datos o fundamentos de las garantías, pues las primeras son explícitas, mientras que las segundas, implícitas. Las garantías no se validan por sí mismas, razón por la cual es necesario que exista un respaldo. En el ejemplo anterior, para poder hacer una aseveración tal como: "Los hombres nacidos en Bermudas son británicos", debemos respaldar dicha afirmación, ya que, como afirma Toulmin, la mayor dificultad dentro del proceso de argumentación yace precisamente en especificar las garantías y fundamentar la naturaleza de su "autoridad". Como se mencionó anteriormente, las garantías son presuposiciones, las cuales deben ser justificadas o respaldadas. El respaldo debe ser considerado cuidadosamente, puesto que su pertinencia se ubica precisamente en el campo o ámbito en el cual el argumento es enunciado.

El problema de las garantías, afirma Toulmin (1958), es que en ocasiones existen varias formas de argumentar en distintas direcciones, cada una con sus respectivas premisas y evidencias. Ante estas circunstancias, se deben sopesar los distintos argumentos en conflicto, de tal suerte que se pueda elegir aquella línea de argumentación que conlleve mayor peso. Ahora bien, las garantías son presuposiciones que no se validan *per se*, sino mediante el respaldo (*backing*) que permita justificarlas. Este respaldo debe ser considerado cuidadosamente, puesto que su pertinencia se ubica precisamente en el campo o ámbito en el cual el argumento es enunciado. Los cuatro puntos mencionados anteriormente, agrega Toulmin, son indispensables para conferirle solidez al argumento.

Una vez que el argumento cuenta con solidez, el siguiente paso consiste en establecer la fuerza del mismo. Toulmin *et al.* (1984) sostienen que los cuatro pasos anteriores son sólo el principio del argumento, puesto que las garantías son de diversa naturaleza, y por

ello pueden conferir diferentes grados de fuerza de conclusión. Es aquí donde entra el segundo nivel de análisis, mismo que los autores denominan como "la fuerza" de un argumento, la cual se encuentra compuesta por los calificadores modales (*qualifiers*), la refutación (*rebuttal*) y las excepciones (*exceptions*). Según Toulmin (1958) y Toulmin *et al.* (1984), esta fuerza se expresa mediante adverbios que denotan una cualidad: *necesariamente, tentativamente, probablemente, aparentemente, quizás, muy probablemente, seguramente*, etc. Por ende, puede que no resulte suficiente especificar los datos, la aserción y la garantía, y que sea necesario añadir un calificador modal. El calificador se encuentra determinado y complementado por las refutaciones y/o por las condiciones de excepción, y se localiza justo frente a la conclusión, pues determina las condiciones mediante las cuales se puede aseverar algo.

Por último, las refutaciones y excepciones ayudan a establecer los límites de la aserción mediante un calificador modal; asimismo, permiten afirmar o rechazar un argumento. Toulmin *et al.* (1984) clasifican las refutaciones en dos tipos: a) cuando los datos, las garantías y el respaldo dan apoyo a la aserción solamente de forma parcial; y b) cuando existe soporte de los datos, garantías y respaldo solamente en ciertas condiciones. En el primer caso, el argumento carecerá de toda su fuerza, mientras que en el segundo, el argumento adquiere validez en la ausencia de condiciones particularmente excepcionales.

Sin lugar a dudas, gracias a la adopción del modelo de argumentación de Toulmin en la validación de los exámenes de alto impacto, se dio un cuantioso salto en la noción de validez y confiabilidad. Asimismo, otra de las ventajas del modelo de Toulmin se debe a que éste no busca evidencia perfecta, sino más bien plausible.

> *If we always waited until absolutely rigorous arguments could be constructed before we acted with reasoned confidence, we would be overtaken by events before we had occasion to act. In practice, therefore, it is often reasonable to base our conclusions on something less than absolutely perfect evidence. We then put forward our claims, not as being formally irrefutable, but rather as being practically strong or reliable* (Toulmin *et al.*, 1984: 81).

Esta noción de verosimilitud *versus* perfección en la evidencia resulta indispensable en el constructo de un argumento de validación, puesto que, como bien se sabe, la suma de los elementos de validez no necesariamente le confiere validez a un examen, sino más bien, la validez se vale de diversos mecanismos para ser plausible. Aunado a lo anterior, el argumento empleado puede irse modificando a lo largo del tiempo en caso de que nuevos hallazgos fundamenten la implementación de cambios en el modelo previamente concebido.

Los enfoques de argumentación aplicados a la evaluación de exámenes

A continuación se presentan los enfoques de argumentación dentro del proceso de validación que mayor influencia han tenido en la evaluación de alto impacto. A lo largo de dos décadas, estos enfoques se han ido desarrollando y retroalimentando entre sí; sin embargo, no fue sino hasta comienzos del siglo XXI que comenzaron a adquirir mayor fuerza y consistencia; por esta razón, hace apenas una década que comenzaron a ser adoptados dentro del proceso de validación por las grandes instituciones internacionales que diseñan y administran exámenes de alto impacto. El orden en el cual se presentan corresponde meramente a un criterio alfabético; posteriormente se discutirán las aportaciones de cada uno de ellos al campo de la evaluación de alto impacto.

El enfoque basado en el argumento de validación de Bachman

Lyle Bachman representa un pilar fundamental dentro del diseño y evaluación de exámenes de lengua. Si bien es cierto que su área de especialización es la evaluación en el área de la lingüística aplicada, su discurso sobre las consecuencias benéficas de la evaluación en programas y mejoras educativas ha contribuido considerablemente a la construcción de los enfoques teóricos basados en la lógica argumentativa y destinada al diseño y evaluación de exámenes de alto impacto. Bachman (2005) reconoce la importancia que juega la validez en relación con las consecuencias del uso de exámenes y enfatiza la necesidad de contar con un enfoque argumentativo que permita llegar a conclusiones o afirmaciones respecto de una forma de evaluación específica, un grupo de examinandos específico y un contexto específico.

A poco más de dos décadas de la publicación de Bachman: *Fundamental Considerations in Language Testing* en 1990, la discusión en torno a la problemática de la evaluación continúa vigente. Después de una década de discusión respecto al tema, Bachman (2005) retoma dos preguntas fundamentales formuladas por Spolsky (1981) en las primeras discusiones en materia de evaluación de lenguas que se llevaban a cabo a principios de la década de los ochenta: "*How sure are you of your decision?*", y "*How sure are you of the evidence that you're using to make that decision?*" (Spolsky, 1981, cit. en Bachman, 2005: 5).

Indudablemente, ambas preguntas denotaban ya la importancia que representaba la toma de decisiones justas y equitativas. Bachman (2005: 5) retoma ambas preguntas y las reformula de la siguiente manera: "*How convincing is the argument for using the assessment in this way?*"; y la segunda: "*How credible is the evidence that supports this argument?*". Para Bachman, un aspecto fundamental en su trabajo de validación se refiere precisamente al uso que se le dará a los exámenes. Su enfoque (2005) retoma la estructura de argumento de validación elaborada por Mislevy *et al.* (2003), la cual se sustenta en el modelo de argumentación de Toulmin. Al enfocar de manera práctica el modelo de Toulmin, y al relacionarlo con la evaluación de la adquisición de nuevas lenguas, Bachman (2005: 9-10) describe sus partes de la siguiente manera:

- Aserción: se refiere directamente a la interpretación que queremos establecer. Es una conclusión sobre lo que el examinando conoce y puede hacer.
- Datos: es la información en la cual se sustenta la aserción; es decir, las respuestas del examinando.
- Garantía: son las proposiciones que empleamos para justificar las inferencias.
- Respaldo: en la evaluación, ésta proviene de la teoría, experiencia previa o evidencia del proceso de validación, así como del análisis de necesidades de la tarea del dominio al cual se encuentra circunscrito el examen.
- Refutación: corresponde a las posibles fuentes de invalidación.
- Evidencia de refutación: consiste en la evidencia que apoye, debilite o rechace una explicación alternativa.

Bachman y Palmer (2010) retoman nuevamente las premisas básicas del enfoque argumentativo de Bachman (2005), pero parten de la equidad en el uso de exámenes; es decir, de las consecuencias del uso de la evaluación por las partes interesadas, que generalmente son las que toman decisiones respecto del desempeño de un candidato. Consecuentemente, la responsabilidad en la toma de decisiones resulta un punto crucial en la validez y la equidad en los exámenes, de tal suerte que quienes diseñan exámenes requieren de herramientas que les permitan rendir cuentas, a las partes interesadas, acerca de las bases en las que se sostienen sus conclusiones respecto

del desempeño de un candidato. Una decisión errónea, tanto en perjuicio como en beneficio de un examinando, no solamente afectará la noción de equidad, sino que además mermará la confianza por parte de los examinandos y de las instituciones. Por tal motivo, una cuestión fundamental será la de proporcionar información suficiente para que quienes toman decisiones, lo hagan a favor de las consecuencias benéficas de las partes interesadas: examinandos, administradores escolares, autoridades, agencias educativas y quienes diseñan los exámenes. Lo anterior se logra mediante el argumento de uso de la evaluación, definido por Bachman (2005) y retomado por Bachman y Palmer (2010) como *assessment use argument (AUA)*.

Dicho argumento proporciona un marco conceptual que vincula el desempeño de un sustentante en un examen con la toma de decisiones y sus consecuencias, por una parte, mientras que por otra, de manera inversa, provee el fundamento y las bases para justificar las decisiones que se toman para el diseño y desarrollo de exámenes. Por consiguiente, el enfoque de Bachman y Palmer (2010), al igual que el de Mislevy *et al*. (2003), cumple dos propósitos: servir como argumento de interpretación y uso, y también como esquema para el diseño y desarrollo de exámenes.

El enfoque basado en el argumento de validación de Kane

Michael Kane es especialista en validez educativa y actualmente ocupa el cargo directivo de validez de exámenes en el Educational Testing Service (ETS), mismo que asumió en 2009. Dicha institución es la mayor organización a nivel mundial dedicada a la investigación y evaluación educativa. Kane (1992) fue el primero en considerar la validación de exámenes como un proceso basado en la argumentación, siguiendo el modelo de argumentación de Toulmin. Esto debido a dos razones: en primer lugar, porque los estándares para la evaluación educativa (AERA, APA y NCME, 1985; 1999) no ofrecían realmente una guía apropiada para establecer el vínculo entre los resultados de los exámenes y la interpretación para hacer uso de ellos eficazmente y de manera imparcial; y en segundo lugar, debido a que el marco de validez ofrecido por Messick (1989) resultaba abstracto y difícil de aplicar. En consecuencia, Kane desarrolló el enfoque basado en la argumentación concebido como "un marco sustentado en la interpretación para recolectar y presentar evidencia válida y explícita asociada con la plausibilidad de varias suposiciones e inferencias involucradas en la interpretación" (Kane, 1992: 2). El enfoque, desde sus inicios, se concibió compuesto por dos tipos de argumentos: el interpretativo y el de validez. El primero contenía el razonamiento que partía de los resultados de un examen hacia las afirmaciones y decisiones que se efectuaban con base en éstos; mientras que el segundo presentaba el caso de plausibilidad, débil o fuerte, del primero.

A lo largo de más de dos décadas, Kane (1992; 2002; 2006; 2013) ha modificado y adaptado su enfoque de argumentación para la validación —*argument-based approach to validity*— de acuerdo a las necesidades que surgen en materia de evaluación. El autor postula que es necesario un argumento mediante el cual se sustente la relación subyacente entre el resultado de un examen y la interpretación que de éste se deriva. "Validar la interpretación del puntaje de un examen es sustentar la plausibilidad del argumento de interpretación correspondiente con su apropiada evidencia" (Kane, 1992: 527).

Dentro del argumento de validez, el autor enfatiza la necesidad de contar con varios tipos de evidencia organizados de manera coherente y convincente, además de su posible refutación. El proceso de argumentación puede ser debilitado mediante la evidencia, pero enmendado en sus fallas y deficiencias; sin embargo, agrega Kane (2013), el mayor problema en la evaluación se debe a suposiciones "ocultas"; es decir, aquéllas que se han

omitido dentro del proceso de interpretación. Evidentemente, cualquier falla dentro del argumento de validez podría ser corregida o planteada de distinta forma, pero ¿qué sucedería si se omitieran suposiciones importantes no por negligencia, sino por desconocimiento de su existencia? Por ello, Kane (2013) afirma que es importante que constantemente se evalúen los argumentos que se postulan, con el fin de que se cerciore la pertinencia de cada uno de ellos.

Desde su concepción inicial, Kane (1992) y Kane *et al.* (1999) ya concebían, dentro del argumento interpretativo, la concatenación de una serie de inferencias que incluyen: generalizaciones, extrapolaciones, predicciones, explicaciones, inferencias basadas en la teoría y decisiones basadas en los puntajes obtenidos en una prueba. Sin embargo, si bien es cierto que a lo largo de dos décadas el enfoque de Kane había girado exclusivamente en torno al argumento de interpretación de los resultados, recientemente este autor introdujo el argumento de uso de un examen propuesto por Bachman (2005) y retomado por Bachman y Palmer (2010), dentro de su enfoque de validación. De ahí que Kane ahora reconozca que "la validez no es una propiedad del examen, sino más bien es una propiedad de las interpretaciones y usos que se proponen de los resultados de un examen" (Kane, 2013: 3). De esta manera, el procedimiento para validar un examen bajo el enfoque basado en la argumentación, consiste en desarrollar argumentos de interpretación y de uso del examen (*interpretation/use argument*) y, posteriormente, validar dichos argumentos.

El argumento de interpretación y de uso

De manera muy general, el enfoque actual de Kane (2013) se divide en dos tipos de inferencias: semánticas y políticas. Las primeras incluyen la inferencia de los resultados observados, de generalización y de extrapolación; en tanto que las segundas exclusivamente engloban la toma de decisiones.

Inferencia de los resultados observados

De acuerdo con Kane (1992; 2002; 2006; 2013), el primer paso necesario para el proceso de argumentación de la evaluación es contar con las evidencias producidas por el alumno: es decir, con las que se obtienen del examen que el candidato ha resuelto. Una vez que se tiene esa muestra, se la evalúa y se lleva a cabo la primera inferencia, la cual consiste en afirmar que el puntaje observado es reflejo del comportamiento observado. En otras palabras, se busca demostrar que los mecanismos de evaluación fueron los apropiados y que éstos fueron aplicados de la manera correcta. Esta inferencia nos lleva al puntaje observado en el desempeño de quien resolvió el examen.

Inferencia de generalización

El segundo tipo de inferencia se refiere a la confiabilidad. En el caso de las pruebas de respuesta seleccionada, se espera que los ítems sean representativos homogéneamente en cuanto al grado de dificultad del universo de conocimientos que se desea evaluar; mientras que en relación con las pruebas de respuesta construida, se espera que las rúbricas estén bien elaboradas, que los correctores sepan emplearlas efectivamente y que, por lo tanto, su evaluación sea consistente. En realidad, la importancia del puntaje se debe a que con base en él se toman decisiones y se busca establecer o afirmar lo que el candidato será capaz de llevar a cabo en contextos más amplios. Kane (2013) destaca que en un examen solamente se puede evaluar una pequeña muestra de desempeño, y no toda una amplia gama de posibilidades existentes en el universo de dominio para el cual es diseñado un examen.

Inferencia de extrapolación

El tercer tipo de inferencia involucra la extrapolación de los resultados obtenidos en la prueba hacia un contexto circunscrito en

el "mundo real"; en este caso, por ejemplo, el educativo. Este tipo de inferencia es crucial, y es una de las más difíciles de considerar, pues implica extrapolar el desempeño de un candidato, de un examen a las situaciones reales. La extrapolación es sumamente compleja e incluye ciertos procesos que van más allá de los análisis estadísticos. En este caso, se busca elaborar aserciones que contemplen situaciones de uso dentro del "mundo real". A este respecto, cabe señalar que Kane (2013) menciona que el argumento que mayor atención amerita es aquél que resulta más débil. En el caso de los exámenes de alto impacto, el argumento más endeble es, justamente, el de la extrapolación.

Kane (2006) menciona dos tipos de evidencia que pueden emplearse para evaluar la inferencia de extrapolación: analítica y empírica. La evidencia analítica se emplea generalmente durante el diseño y constructo del examen, mientras que la empírica busca relacionar los puntajes observados en un examen con otros tipos de muestras o de puntajes asociados con el dominio. Recordemos que el fin último de los exámenes es tomar decisiones, por lo cual la inferencia de extrapolación resulta indispensable dentro del argumento de interpretación.

El mayor problema, como se describió anteriormente, radica en qué tan estandarizado se encuentre el examen, pues nuestra confianza en la inferencia de extrapolación dependerá de la relación entre el examen y el dominio meta; de tal suerte que aquellos exámenes que cubran un amplio espectro del dominio en cuestión podrán considerarse plausibles y, por ende, la inferencia de extrapolación no requerirá de demasiado soporte empírico. Sin embargo, en caso contrario, cuando las tareas de un examen difieran sustancialmente del dominio meta, el respaldo de la extrapolación será imprescindible. Si el argumento que se establece a partir de las inferencias de interpretación y de uso es validado mediante los mecanismos descritos anteriormente, poseerá una fuerte presuposición en favor de éste, con lo que se podrá justificar el uso de los puntajes obtenidos en un examen.

Inferencia de toma de decisiones

La cuarta y última inferencia implica la toma de decisiones. Ésta es la única inferencia que Kane (2013) considera de tipo político, dado que involucra ciertas normas previamente establecidas por la autoridad en cuestión (universidades, empresas, gobierno, entre otras), respecto de qué sustentante es apto para hacer uso de las habilidades/competencias demostradas en diferentes contextos: académicos, profesionales o con fines migratorios. Kane (2013) retoma la postura de Bachman y Palmer (2010) y afirma que en la toma de decisiones deberá establecerse un balance de las consecuencias positivas y negativas de las mismas, optando por aquéllas en donde las positivas sobrepasen a las negativas.

El modelo de validación de Mislevy

Robert Mislevy es especialista en estadística y medición educativa. Al igual que Kane, forma parte del ETS, en donde dirige el área de estadística y medición desde 2010. Retomando la noción de argumentación de Toulmin (1958), Mislevy *et al.* (2003) desarrollaron el modelo de evaluación conocido como *evidence-centered design* para explicar el papel central que juega el razonamiento probatorio a través de la evidencia. Dicho modelo partió de la necesidad de establecer un argumento que permitiera relacionar lo que los estudiantes dicen o hacen en un examen con lo que hacen en contextos más amplios.

Según Mislevy *et al.* (2003), los pasos para el diseño de exámenes son equivalentes a los que se deben seguir en el proceso de validación de los mismos. Los autores incluyen cuatro pasos fundamentales, cada uno de los cuales está compuesto, a su vez, por diversos componentes: el análisis del dominio, el modelado del dominio, el marco conceptual de evaluación

y la arquitectura de la prueba, la cual engloba la aplicación y puntuación de la tarea (Mislevy, 2006; 2009; Mislevy *et al*., 2003).

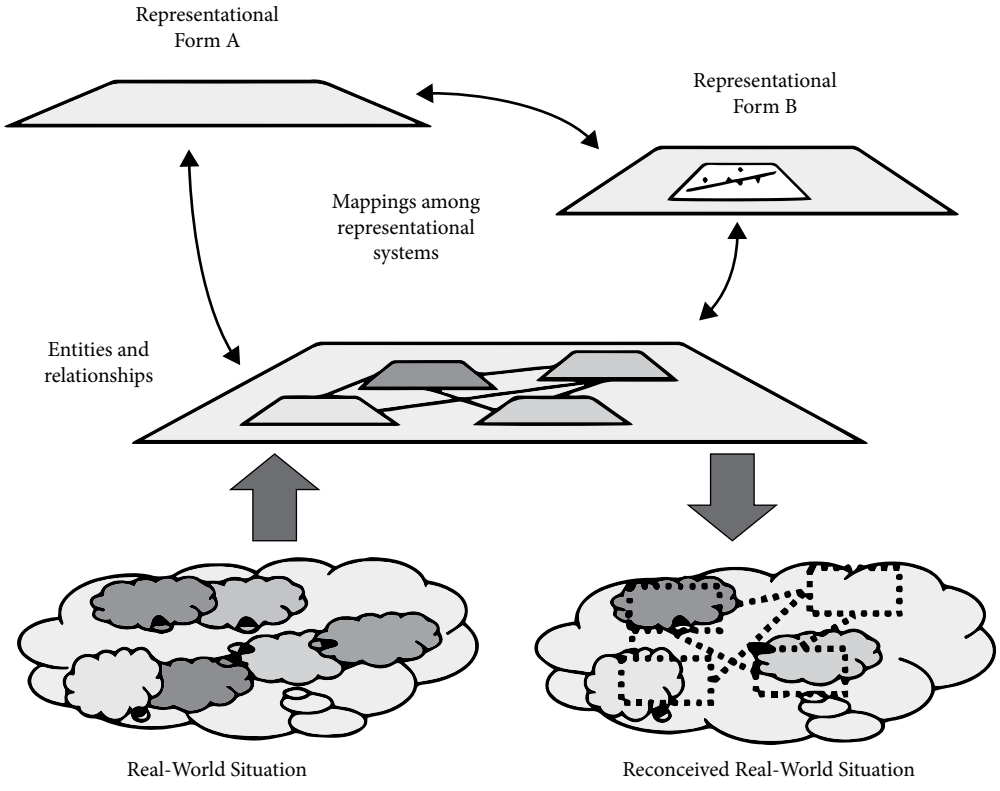
De manera sucinta se describe cada uno de los componentes del diseño centrado en la evidencia (*evidence-centered design*): en primer lugar, es importante hacer un análisis del dominio, en otras palabras, de aquellas situaciones de la "vida real" en las cuales deseamos realizar la evaluación. En segundo término, los autores describen el modelado del dominio, el cual se conforma de las tareas y de todos los instrumentos que se requerirán para evaluarlas; por ejemplo, las rúbricas. En tercer lugar está el marco conceptual de evaluación, en donde se encuentran todos los procesos necesarios para determinar la confiabilidad de una prueba, llámense especificaciones, requerimientos operativos, modelos estadísticos o detalles de rúbricas, entre otros. Generalmente, aunque no siempre, dentro de este marco se encuentran todos los análisis estadísticos que se requieren y que son indispensables para determinar si los instrumentos que han sido diseñados, así como los mecanismos para medirlos, cumplen con la función para la cual fueron creados. Finalmente, la cuarta sección es la arquitectura del examen, misma que incluye la evaluación operativa; en ésta se considera el resultado de los exámenes y las partes involucradas: examinadores, administradores, examinandos, así como aquellos que toman las decisiones.

El modelo de Mislevy es bastante complejo y a través de los años ha sido reelaborado por el autor (2006; 2009); no obstante, los puntos descritos anteriormente constituyen su base fundamental. En la actualidad, dicho modelo es conocido como *model-based reasoning*, o modelo basado en el razonamiento, con sustento en una perspectiva sociocognitiva (Mislevy, 2009). Para Mislevy, un elemento esencial de la validez de una prueba resulta si al emplear un modelo subyace una base sólida para la organización de las observaciones y para la guía de las acciones que deberán considerarse en las situaciones para las cuales el examen fue destinado.

Mislevy concibe su modelo en tres planos (Fig. 2); en la parte inferior se muestran diversas manifestaciones del mundo real. De acuerdo con este autor (2006; 2009), el sustento teórico y empírico es imprescindible para poder diseñar las tareas mediante las cuales se evaluará a un estudiante; de ahí la importancia del constructo dentro de la línea de validez de argumentación. Ahora bien, para acceder del plano inferior al intermedio se deben considerar previamente los fundamentos teóricos y empíricos de los procedimientos necesarios para evaluar las tareas.

En el plano intermedio, denominado por Mislevy como "plano semántico", se muestran las entidades, relaciones y propiedades del modelo; es decir, la serie de argumentos que tendrían que llevarse a cabo —a manera de filtro— para hacer inferencias y tomar decisiones respecto de las habilidades de un estudiante. Se trata, entonces, de un modelo basado en la teoría y la experiencia —el constructo de la prueba— que le da sustento científico a esta capa del modelo. Según Mislevy (2009) este plano intermedio, conformado por modelos teóricos, funge como intermediario entre el complejo "mundo real" (plano inferior) y el abstracto y rígido "mundo estadístico" que se emplea para validar exámenes (plano superior).

Figura 2. El proceso de concebir nuevamente una situación del mundo real a través de un modelo



Fuente: tomado de Mislevy, 2009: 12.

En el plano superior figuran los sistemas de medición educativos que darán soporte al razonamiento y argumento previamente definidos. Mislevy lo esquematiza a manera de fórmulas y ecuaciones que representan el proceso estadístico que debe efectuarse para validar un examen y, así, a partir de sus resultados, poder tomar decisiones. El analista, a través del modelo de razonamiento, conecta nuevamente el sistema científico de validez de la prueba con el "mundo real", lo cual implicaría la toma de decisiones. Consecuentemente, se parte de un examen concebido dentro del mundo real y el objetivo final es, justamente, la inferencia de lo que un sustentante será capaz de realizar en un contexto determinado para el cual está siendo examinado.

Discusión sobre los enfoques de argumentación para exámenes

Los tres enfoques descritos anteriormente han aportado cuantiosamente al proceso de validación de exámenes. El modelo de argumentación de Mislevy es por demás complejo y hace gran énfasis en los modelos matemáticos que deberán utilizarse para lograr la confiabilidad de la prueba, lo cual dificulta su aplicación y deja de lado inferencias importantes que deben ser tomadas en cuenta al momento de validar un examen. Otra de las desventajas que el mismo Mislevy reconoce (2006) respecto de los modelos es que éstos no son perfectos y resulta difícil adaptarlos, ya que no siempre es claro definir cuáles podrían

ser los criterios para flexibilizarlo al momento de emplearlo.

La mayor aportación de Bachman (2005) y Bachman y Palmer (2010) se debe, desde luego, al denotado énfasis en las consecuencias de los exámenes; de hecho, el argumento de validación busca como fin último el uso que se hará de los exámenes. La crítica de Bachman y Palmer (2010) hacia las propuestas de Kane y Mislevy reside fundamentalmente en que estos autores no incorporan el concepto de uso (Kane, 2006) o lo hacen de manera tangencial (Mislevy, 2006). Por ello, Kane (2013) incorpora dentro de su discurso el argumento de interpretación y uso de los exámenes. No obstante, a diferencia del peso mayoritario que le confieren Bachman (2005) y Bachman y Palmer (2010), Kane afirma que ambos argumentos —de interpretación y de uso— son igual de importantes. Pese a lo anterior, la línea discursiva de Bachman (2005) y Bachman y Palmer (2010) hace referencia profusa a las consecuencias y el uso de los exámenes, tomando en cuenta las nociones de equidad y de ética, pero dejando de lado la importancia de las otras inferencias que son fundamentales para crear un argumento sólido en cada uno de los estadios que se requiere para llegar de los datos a la aserción.

El enfoque de Kane, por otra parte, es el que mayor desarrollo, profusión y aplicaciones ha tenido desde la década de los noventa, cuando incorporó por vez primera el argumento por el cual Cronbach (1988) y Messick (1989) abogaban por concebir la validez de forma unitaria. Su enfoque, además, incorporó desde sus inicios el modelo argumentativo de Toulmin (1958) para la validación de exámenes. Como bien señala Kane (2013), una de las mayores ventajas de su enfoque se debe a que es flexible y moldeable a la situación para la cual se pretenda validar un examen. Las inferencias que el autor propone no resultan a discreción del experto en evaluación, sino más bien se definen de acuerdo con las necesidades para las cuales se pretende evaluar. Por ello, el autor describe detalladamente en qué casos será conveniente incluir cada una de las inferencias que él propone, así como la evidencia que sería necesaria para fundamentarlas.

Otra de las ventajas del enfoque de Kane se debe precisamente a que, al igual que Toulmin, su marco se encuentra sujeto a ser refutado mediante la contraevidencia. Sin lugar a dudas esto representa una de las grandes aportaciones del enfoque, pues precisamente el objetivo de la validez no es aceptar y aprobar un modelo de elaboración/evaluación de exámenes, sino encontrar fallas y huecos que permitan su perfeccionamiento, desarrollo y evolución. Por esta razón, el enfoque de Kane se ha adaptado a las condiciones que imperan en materia de evaluación, así como de validez de los exámenes.

Finalmente, el reconocimiento de Kane (2013) respecto del uso y las consecuencias de los exámenes, además de la interpretación de sus resultados, le confiere a su enfoque la solidez que se requiere cuando se trata de exámenes de alto impacto, mismos que conllevan importantes implicaciones en la vida de los examinandos. Asimismo, implica un reconocimiento de la importancia que actualmente juegan la ética y la equidad en la toma de decisiones y, por lo tanto, la flexibilidad, así como la capacidad de adaptación y evolución de su enfoque.

Aplicación del argumento de validación a los exámenes de alto impacto

Los alumnos no hispanohablantes que desean ingresar al posgrado de la UNAM deben aprobar un examen de posesión de la lengua española que es administrado por el Centro de Enseñanza para Extranjeros (CEPE). El examen actual consta de cuatro pruebas objetivas: comprensión auditiva, comprensión de lectura y vocabulario, y estructuras; y una prueba objetiva: expresión oral. Sin embargo, no se evalúa la expresión escrita, habilidad imprescindible en el ámbito académico. Por

ello, el CEPE se dio a la tarea de diseñar un examen destinado específicamente a aquellos aspirantes no hispanohablantes que pretenden cursar estudios de grado o de posgrado en la UNAM. El examen aún se encuentra en fase de pilotaje, pero se espera que pronto pueda emplearse como nuevo criterio que avale los conocimientos lingüísticos de los aspirantes no hispanohablantes, e incluso de los propios aspirantes nativo-hablantes a cursar estudios universitarios. Sin embargo, debido a la dificultad que subyace a la implementación y validación de pruebas de respuesta construida, como la escrita, Weigle (2002) enfatiza las múltiples consideraciones que deberán tomarse en cuenta para el diseño de éstas.

Por ello, siguiendo la lógica argumentativa propuesta por Toulmin (1958) y su aplicación al enfoque para validar exámenes propuesto por Kane (2006), se presenta un ejemplo para validar la prueba escrita del examen de español con fines académicos (EXELEAA) que diseñó el CEPE de la UNAM (Fig. 3).

De éste, se desprende la siguiente aserción:

> La prueba escrita del examen de español académico (EXELEAA) mide eficazmente la habilidad del examinando para utilizar el lenguaje escrito requerido a nivel posgrado. Asimismo, la puntuación es útil y clara para que quienes se encargan de tomar decisiones lo hagan de manera justa y equitativa.

Para poder validar esa aserción, se plantean —de acuerdo con el enfoque de Kane— cuatro inferencias que deberán sustentarse en evidencia para poder llevar a cabo una interpretación. Estas cuatro inferencias son: evaluación, generalización, extrapolación y utilización. Sin embargo, el autor sugiere que cuando se trata de exámenes de alto impacto en donde la toma de decisiones podría afectar decisivamente la vida de los examinandos se incluyan dos inferencias más: la descripción del dominio y la explicación de los resultados (Kane, 2006). A continuación se describen estas inferencias adaptadas a la prueba del examen de español para el ámbito académico.[3]

Primera inferencia: el argumento de la descripción del dominio

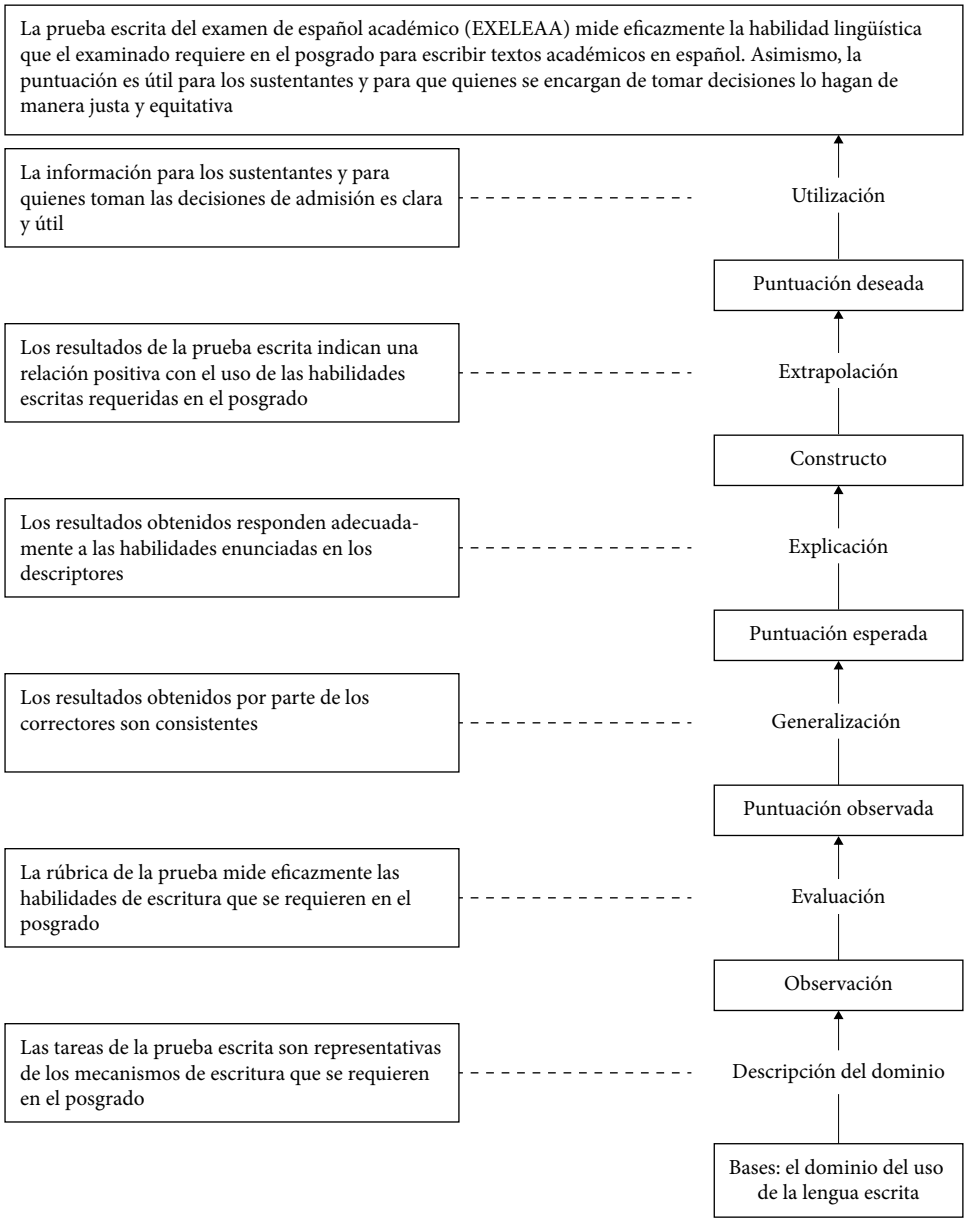
De acuerdo con Kane (2006), resulta imprescindible describir cuidadosamente el ámbito en el cual se pretende evaluar; en ámbitos tan amplios como el académico, es importante que las tareas que han sido diseñadas sean representativas del universo para el cual se pretende evaluar. La prueba escrita fue diseñada partiendo de los requerimientos de escritura de los estudiantes universitarios; no obstante, los mecanismos de evaluación varían considerablemente de una disciplina a otra y también de un grado a otro, es decir, será necesario investigar en un estudio exploratorio cuáles son los mecanismos de evaluación de los diversos posgrados de la UNAM.

Segunda inferencia: el argumento de la evaluación

Una vez que se administra el examen, se recaban los datos con el fin de medirlos. Este proceso es conocido como evaluación. En el caso aquí ejemplificado, el procedimiento consiste en obtener las muestras de la prueba escrita y corregirlas para obtener un puntaje. Dentro de este proceso, resulta fundamental contar con un instrumento que mida lo que se pretende medir. Estos instrumentos comúnmente son denominados "rúbricas", o "escalas de medición". En el presente caso se trata de una rúbrica analítica que da cuenta detallada de los diversos aspectos que se desean evaluar en los escritos. Ahora bien, existen diversos mecanismos para garantizar la eficiencia de las rúbricas. En este caso se emplearán dos mecanismos complementarios: en primer lugar, se identificará cada uno de los puntos que se miden en la rúbrica y se correlacionarán con las habilidades de escritura académica que son necesarias para ser competente

3 El proceso de validación de la prueba se encuentra aún en proceso.

Figura 3. Argumento de validación para la prueba escrita del examen de español académico (EXELEAA) del CEPE, UNAM



Fuente: elaboración propia.

académicamente en el posgrado (Rosenfeld *et al*., 2004). Una vez que se identifiquen los alcances y limitaciones de la rúbrica, se llevará a cabo un proceso estadístico, tras la evaluación de la prueba, con los resultados obtenidos en cada uno de los rubros previamente definidos en la rúbrica. El fin de este estudio estadístico será distinguir los componentes de la rúbrica que funcionan adecuadamente, de los que no lo hacen.

En caso de ser necesario, se llevarán a cabo las modificaciones necesarias para afinar la

rúbrica y se podrán incorporar nuevas tareas que permitan incluir otras habilidades de escritura necesarias en el posgrado. Finalmente, si determinadas habilidades no pudieran ser medidas ni incluidas en la prueba escrita, se sugerirán medios alternos que les permitan a las diversas facultades garantizar que el candidato a estudios de posgrado cuenta con dichas habilidades.

Tercera inferencia:
el argumento de generalización

La inferencia de generalización permite demostrar que los resultados obtenidos en una muestra son consistentes y, por lo tanto, generalizables y confiables. Esto quiere decir que si un candidato presentara varias veces diversas modalidades de la prueba escrita, pero en condiciones similares de aplicación y sin variaciones en su conocimiento a evaluar, su resultado sería el mismo o muy similar. Lo anterior querría decir que las diferentes versiones de la misma prueba son equiparables entre sí, y que el puntaje otorgado por el mismo corrector, y entre correctores, es consistente. En el caso de las pruebas de respuesta construida, éstas generalmente son evaluadas por dos correctores, y cuando existe discrepancia de más de 10 por ciento en el resultado final, se revisa por un tercer corrector experimentado. En caso de no haber consistencia en las evaluaciones, se tendrán que llevar a cabo talleres de capacitación y retroalimentación para los evaluadores.

Cuarta inferencia:
el argumento de explicación

Kane (2006) afirma que el argumento de validez puede prescindir de la explicación, pero sugiere que en los casos en los cuales uno de los componentes de inferencia sea la interpretación de resultados, como es el caso de la evaluación de alto impacto, debería existir una inferencia fundamentada en la explicación. En este caso, los descriptores del nivel de actuación de los examinandos deberán estar sustentados de manera sólida en la teoría lingüística que respalda al examen. Asimismo, los puntos de corte de los diferentes niveles de desempeño deberán encontrarse claramente definidos y deberán correlacionar positivamente con el puntaje demostrado por los candidatos.

Quinta inferencia:
el argumento de extrapolación

En el caso de los exámenes de alto impacto, una de las mayores dificultades que se ha identificado radica justamente en lo difícil que resulta extrapolar los resultados del examen al "mundo real". Lo anterior se debe a que normalmente el universo para el cual se evalúa es demasiado amplio, como es el caso de la escritura académica. Sin embargo, Kane (2013) menciona que el mayor énfasis en la validación de la argumentación deberá encontrarse justamente en el argumento más endeble.

Dado que la prueba escrita se encuentra aún en fase de pilotaje, se optó por trabajar con estudiantes de posgrado no hispanohablantes ya inscritos en la UNAM. Para poder llevar a cabo la extrapolación, se diseñó un estudio cualitativo basado en entrevistas semiestructuradas que permitieran dar cuenta de las percepciones y dificultades que encaran los estudiantes al momento de escribir textos académicos o exámenes escritos en el posgrado. Aunado a lo anterior, se adaptó el cuestionario diseñado por Rosenfeld *et al.* (2004), en aras de dar cuenta del dominio que deben poseer los estudiantes para escribir dentro del ámbito académico. La triangulación de estos dos instrumentos con el puntaje observado en la prueba permitirá determinar si, efectivamente, existe una correlación positiva entre obtener un buen puntaje en la prueba escrita y ser académicamente competente al momento de escribir en el contexto universitario.

Sexta inferencia:
el argumento de toma de decisiones

La inferencia respecto de la toma de decisiones es responsabilidad de quienes utilizan el

examen; por ello, es menester que los encargados de diseñar exámenes de alto impacto elaboren documentos que garanticen que los resultados de los exámenes puedan ser interpretados correctamente por las partes interesadas. En el caso aquí planteado, debido a que no se cuenta actualmente con una guía que contenga los descriptores del nivel de actuación del candidato, ni de ejemplos de actuación en cada uno de los descriptores, se elaborará dicho documento y también se diseñará un cuestionario que permita dar cuenta de su utilidad para las partes interesadas. Esto les permitirá a las facultades de la UNAM tomar la mejor decisión de acuerdo con los requerimientos de escritura que se requieren en cada uno de sus posgrados.

Conclusiones

A partir de la segunda mitad del siglo XX, la evaluación adquirió un carácter científico y la validez de constructo constituyó el eje rector para garantizar la fiabilidad de los exámenes. No obstante, debido a las dificultades que se presentaron a finales del siglo XX relacionadas con la materialización de los modelos de validez en la evaluación, incorporada de manera fragmentada y desarticulada, desde inicios del siglo XXI se ha adoptado la argumentación en la validez vista de manera integradora a través de la lógica argumentativa de Toulmin. Así pues, el establecimiento de una aserción respecto del desempeño de un candidato requiere de una serie de inferencias y su respectiva evidencia, con el fin de garantizar la confiabilidad y validez de un examen. Los tres enfoques descritos demuestran el considerable esfuerzo que se ha llevado a cabo durante la última década con miras a garantizar una evaluación justa y equitativa de los exámenes, particularmente en aquéllos de alto impacto, puesto que conllevan importantes implicaciones en la vida de los examinandos. Dentro de estos enfoques destaca el de Kane, mismo que, gracias a su flexibilidad, ha ido evolucionando a lo largo de casi dos décadas desde su concepción inicial, y se ha ido adaptando a las nuevas exigencias e interrogantes que surgen en materia de evaluación. Finalmente, el proceso de validación de la prueba de expresión escrita del CEPE da cuenta de la aplicabilidad del enfoque teórico de la validación a través de la argumentación. Sin lugar a dudas, la concepción de la evaluación vista desde la óptica de la lógica argumentativa resulta una herramienta útil para todos a aquéllos que diseñan y elaboran exámenes de alto impacto.

Referencias

American Educational Research Association/American Psychological Association/National Council on Measurement in Education (1985), *Standards for Educational and Psychological Testing*, Washington, D.C., American Educational Research Association.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education (1999), *Standards for Educational and Psychological testing*, Washington, D.C., American Educational Research Association.

Bachman, Lyle (2005), "Building and Supporting a Case for Test Use", *Language Assessment Quarterly*, vol. 2, núm. 1, pp. 1-34.

Bachman, Lyle y Adrian Palmer (2010), *Language Assessment in Practice: Developing language assessments and justifying their use in the real world*, Oxford, Oxford University Press.

Cronbach, Lee (1988), "Five Perspectives on Validity Argument", en Howard Wainer y Henry Braun (eds.), *Test Validity*, Hillsdale, NJ, Lawrence Erlbaum, pp. 3-17.

Cronbach, Lee y Paul Meehl (1955), "Construct Validity in Psychological Tests", *Psychological Bulletin*, vol. 52, núm. 4, pp. 281-302.

Kane, Michael (1992), "An Argument-Based Approach to Validation", *Psychological Bulletin*, vol. 112, núm 3, pp. 527-535.

Kane, Michael (2002), "Validating High-Stakes Testing Programs", *Educational Measurement: Issues and Practice*, vol. 21, núm. 1, pp. 31-41.

Kane, Michael (2006), "Validation", en Robert Brennan (ed.), *Educational Measurement*, Westport, American Council on Education/Praeger, pp. 17-64.

Kane, Michael (2013), "Validating the Interpretations and Uses of Test Scores", *Journal of Educational Measurement*, vol. 50, núm.1, pp. 1-73.

Kane, Michael, Terence Crooks y Allan Cohen (1999), "Validating Measures of Performance", *Educational Measurement: Issues and Practice*, vol. 18, núm. 2, pp. 5-17.

Messick, Samuel (1989), "Validity", en Robert Linn (ed.), *Educational Measurement*, Nueva York, American Council on Education/Macmillan, pp. 13-103.

Messick, Samuel (1995), "Validity of Psychological Assessment", *American Psychologist*, vol. 50, núm. 9, pp. 741-749.

Mislevy, Robert (2006), "Cognitive Psychology and Educational Assessment", en Robert Brennan (ed.), *Educational Measurement*, Westport, American Council on Education/Praeger, pp. 257-305.

Mislevy, Robert (2009),"Validity from the Perspective of Model-based Reasoning", *Cresst Report*, núm. 752, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, pp. 1-28.

Mislevy, Robert, Linda Steinberg y Rusell Almond (2003), "On the Structure of Educational Assessments", *CSE Technical Report 597*, Los Angeles, pp. 1-67.

Rosenfeld, Michael, Rosaela Courtney y Mary Fowles (2004), "Identifying the Writing Tasks Important for Academic Success at the Undergraduate and Graduate Levels", *Educational Testing Service Research*, Report 42, Princeton, Educational Testing Service.

Toulmin, Stephen (1958), *The Uses of Argument*, Cambridge, Cambridge University Press.

Toulmin, Stephen, Richard Rieke y Allan Janik (1979), *An Introduction to Reasoning*, Nueva York, Macmillan.

Toulmin, Stephen, Richard Rieke y Allan Janik (1984), *An Introduction to Reasoning*, Nueva York/Londres, Macmillan Publishing, Co. Inc./Collier Macmillan Publisher.

Weigle, Sarah (2002), *Assessing Writing*, Cambridge (UK), Cambridge University Press.