Capítulo 8

EL RETO DEL ESTABLECIMIENTO DE ESTÁNDARES DE EVALUACIÓN Y PUNTOS DE CORTE EN EDUCACIÓN SUPERIOR

Mildred López, Gabriela González

INTRODUCCIÓN

La alineación entre el currículo educativo diseñado, los contenidos abordados en el proceso enseñanza-aprendizaje, y la evaluación determina la percepción de que el proceso ha sido efectivo y justo. La evaluación es sin duda uno de los momentos más críticos en la vida estudiantil de aprendices y docentes, es la *hora de la verdad*, el momento donde se demuestra la efectividad del aprendizaje. Si el proceso de aprendizaje ha sido un círculo virtuoso, este es, en sí mismo, un momento de aprendizaje en el que el estudiante recibe evidencias de su desarrollo y sugerencias de áreas o estrategias que aún necesita trabajar. Cuando este proceso deja mucho que desear, el momento está cargado de emociones como miedo, frustración y enojo.

Si bien, la primera capa de análisis es importante, es decir el análisis en lo individual por parte de los estudiantes, para reflexionar sobre qué puedo hacer mejor; el análisis que debemos hacer como educadores y diseñadores de programas también tiene que estar presente. Este análisis exige que la evaluación sea por diseño de alta calidad, que garantice el egreso de estudiantes con el más alto estándar, y que los instrumentos y procesos que utiliza para clasificar a los estudiantes, hayan sido conceptualizados con el mismo nivel de exigencia.

Los objetivos de este capítulo son:

- Analizar los conceptos de establecimiento de estándares en evaluación y puntos de corte.
- Contrastar los principales criterios que se utilizan para el establecimiento de puntos de corte en la definición de estándares.
- Discutir las implicaciones para el aprendizaje a distancia resultado de la pandemia del nuevo coronavirus.

DESARROLLO DEL TEMA

Una definición quizás un tanto simplista del concepto de evaluación es que esta fase del proceso enseñanza-aprendizaje colecta evidencia del desempeño del estudiante. Es posible describir un momento de evaluación como un instante en el tiempo en el que se toma una fotografía al desempeño de una persona. Esta fotografía puede ser tan favorecedora u horrible tanto como el escenario, la iluminación y la escenografía hayan sido dispuestas. También tiene mucho que ver la experiencia del modelo para saber cómo posar y del fotógrafo para gestionar esos elementos en el plató. Siguiendo con la alegoría de las fotografías, para conocer verdaderamente qué es capaz de hacer el *modelo* en la imagen, es importante tener acceso a la mayor cantidad de fotografías posibles donde demuestre su desempeño en una amplitud de contextos, estilos y actitudes. En la evaluación del aprendizaje, es igual. Solo que las fotografías son en realidad los resultados de distintos instrumentos y pruebas que se obtienen en diferentes etapas del desarrollo profesional. En la medida que se tienen más *fotografías*, más se tiene certeza de que la valoración realizada fue correcta.

Boud (2006) afirma que la evaluación genera más ansiedad en los estudiantes e irritación en los docentes que cualquier otro elemento de la educación superior. Para los estudiantes significa quizás horas de estudio adicional el fin de semana, tratar de memorizar cientos de páginas de libros de texto y diapositivas de una presentación. Su preocupación gira al menos, en torno a dos elementos: el primero se refiere al momento donde estarán en el examen o la prueba en la que tienen un tiempo limitado para responder preguntas, y el segundo al juicio que se hará sobre su desempeño con este instrumento. Para los docentes, esta ansiedad se encuentra en el diseño de las pruebas, las fechas límite para entregar reactivos y el estrés de enfrentarse a estudiantes decepcionados o iracundos una vez que se entreguen los resultados (Joughin, 2008).

¿Qué es el establecimiento de estándares?

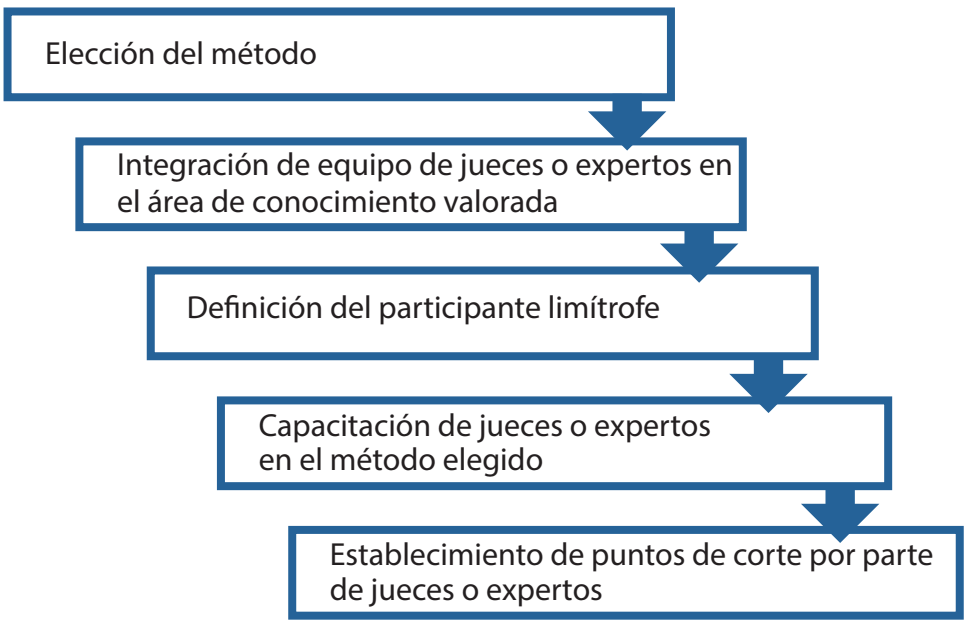
Cuando la evaluación se utiliza además para la toma de decisiones, diversas inferencias surgen a partir de las calificaciones o la respuesta de los evaluados. Estas inferencias tienen grandes repercusiones ya que determinan por ejemplo qué desempeño es *suficientemente bueno* para ser admitido a una universidad o cuánto es suficiente saber de matemáticas para avanzar al siguiente año escolar. Una definición sencilla sobre el establecimiento de estándares es que se refiere al proceso de decidir qué es lo *suficientemente* bueno (Cusimano, 1996). Barman (2008) añade a la discusión diciendo que el establecimiento de estándares es una actividad de generación de políticas, en un nivel más conceptual fija un ideal, mientras que la operacionalización de este ideal es la puntuación de corte descrita en dicha política. Para fines prácticos, el reto en la definición de un estándar se refiere en realidad a la definición de un punto de corte, es decir el nivel mínimo aprobatorio que separa entre desempeño competente del que no lo es.

En modelos educativos basados en competencias que describen un perfil de egreso, la promesa es que cada graduado domina todas estas competencias en un nivel aceptable. Cada competencia es evaluada usando el mismo criterio para todas y todos los graduados de la misma generación y es muy similar a las generaciones anteriores. Esto quiere decir que los

resultados de las evaluaciones deben separar en dos grupos, excluyentes entre sí, aquellos que se conforman al estándar definido, que tienen un desempeño competente y los que no, quienes aún deben trabajar para alcanzar un nivel de desempeño aceptable. Esta evaluación basada en criterios valora el desempeño del participante en sí mismo, independientemente del desempeño de otros miembros del grupo. Es decir que todos los estudiantes que cumplan con los criterios mínimos obtendrán un documento que avala su competencia, ya sea un título o diploma de grado.

Aunque esta definición de absolutos parece intuitiva de entender, cualquier persona que haya puesto un pie en un salón de clases y haya interactuado con estudiantes sabe que una evaluación absoluta, única y objetiva en un solo momento de interacción es imposible de obtener. A decir verdad, el desempeño de los estudiantes varía de ser razonablemente competente, competente acompañado de un docente, o muy competente, por nombrar algunos. Este problema se agrava al traducir un desempeño a una calificación, ¿cómo se compara la competencia de un estudiante que obtuvo un 80% con el que obtuvo un 90% en una escala donde el 70% es el corte aceptable para definir el pase en esta prueba? Muchas de estas pequeñas variaciones son subjetivas y difícilmente traducible a criterios observables.

Figura 1: Pasos para establecimiento de estándares



Otro caso de definición de estándares que frecuentemente enfrentamos en educación superior, son aquellos estándares en los que el desempeño de un participante será comparado con otros que sustentaron evaluación al mismo momento, o con algunos que lo han tomado con anterioridad. Por ejemplo, en este caso, no todos los participantes que sustentan un examen o se esforzaron *mucho* en un programa educativo, tendrán acceso a una beca, o una plaza de

posgrado. He ahí el compromiso de que estas definiciones de estándares y las mismas valoraciones de la prueba tengan una definición, conceptual o empírica, avalada por expertos y un análisis de su validez.

Una adecuada definición de estándares debería ser sensible a distinguir el desempeño de los participantes, tener sustento estadístico, ser confiable y fácil de implementar y aplicar por cualquier persona (Barman, 2008). En los últimos años, diversos autores han generado alternativas para desarrollar estos estándares y sustentar las decisiones que estas conllevan, estos varían en cómo deben llevarse a cabo estos juicios de valor, quién debe participar en emitir los juicios, y en qué información deben basarse.

Sin importar el método utilizado para fijar estos estándares y establecimiento de puntos de corte, el proceso debe contener al menos seis pasos (Ruano et al., 2018), descritos en la Figura 1.

La selección del método depende de los siguientes elementos (Ruano et al., 2018):

- Experiencia previa con el método: utilizar un método con el que se está familiarizado reduce la cantidad de tiempo utilizado en los pasos de la Figura 1, además que el equipo de diseño se siente confiado del proceso que está llevando a cabo.
- Tiempo disponible para la definición de estándares: impacta en la elección de métodos centrados en la prueba que pueden ser valorados por un grupo de jueces o expertos considerando un grupo hipotético, o si es posible realizar un piloto para tener los resultados reales de participantes para los cuales fue diseñada la prueba.
- Tipo de preguntas utilizadas: algunos métodos son recomendados para pruebas con reactivos de opción múltiple, otros de respuestas dicotómicas o preguntas abiertas.
- Validez del método: diversos autores han levantado inquietudes respecto a la validez de las decisiones realizadas por distintos métodos. Por ejemplo, existen cuestionamientos sobre el método de grupos contrastados y el Angoff, las cuales son abordadas más adelante.
- Además de considerar la definición de estándares desde el punto de vista metodológico, es importante que el desarrollo de la evaluación sea analizado como algo integral en el que los resultados sean útiles, y defendibles ante los estudiantes, profesores, desarrolladores de la prueba y demás grupos de interés (Cetin y Gelbal, 2013).

Principales métodos para fijar estándares

Históricamente, han existido numerosas propuestas para fijar dichos estándares, tan solo en 1986 existían cerca de 38 métodos para definir estándares, diez años más tarde en 1996 existían al menos 50 (Cusimano, 1996). De acuerdo con el propósito de la evaluación, los métodos para fijar estándares se pueden clasificar en dos tipos, aquellos centrados en la prueba y los centrados en la persona. El primero se enfoca en que los expertos provean un estimado basándose en los reactivos de una evaluación. El segundo se enfoca en la observación del desempeño de los grupos, por lo que debe tener revisiones a lo largo del tiempo (Downing et al., 2010). La Tabla 1 presenta un vistazo a los principales métodos de acuerdo con esta categorización.

Tabla 1: Métodos centrados en la prueba vs en la persona

Enfoque	Estándar
Centrados en la prueba	Angoff.
	Nedelsky.
	Bookmark o de marcador.
	Ebel.
Centrados en la persona	Grupos contrastados.
	Borderline o métodos de frontera.

En los métodos basados en los criterios de la prueba, el contenido del examen es revisado por un grupo de jueces expertos, entre los que destacan el método de Angoff, el Nedelsky, el método de marcador y Ebel. Mientras que en los métodos basados en la persona se concentran en evaluar el desempeño del sujeto, algunos métodos son el de grupos contrastados y los métodos de frontera. Aunque existen diferentes métodos y variaciones de estos, cada método tiene sus ventajas y desventajas lo que los hace más adecuados para una u otra aplicación (Cetin y Gelbal, 2013). A continuación, se describen algunas de las más utilizadas.

MÉTODO DE ANGOFF

Este es uno de los métodos de evaluación estándar más populares y tiene una larga historia de éxito debido a que las calificaciones son fácilmente obtenidas, los cálculos son simples y el método es fácil de comunicar. Su propuesta nace en los años 70's, Angoff presentó un método sistemático para decidir la calificación mínima para aprobar una evaluación. El método sugiere que previo a aplicar una prueba, se lleve a cabo una revisión de todas las preguntas con un grupo de jueces expertos en el área, llamados panelistas, como lo pueden ser los mismos profesores, o invitados externos y sin ninguna relación con la institución (Papageorgiou y Tannenbaum, 2016).

En esta evaluación, los jueces establecen un puntaje entre el 0 y 100% dependiendo de la dificultad de la pregunta. Se le da un puntaje cercano a 100% a aquellas preguntas en la evaluación que se consideran que una persona calificada podría contestar correctamente y cercano a 0 a las cuales se consideran que el evaluado contestaría de manera incorrecta. La suma de todos los puntos obtenidos es considerada el punto de corte (Zieky, 2001). En la Tabla 2 se presenta un ejemplo donde se tienen 4 panelistas que gradúan la dificultad de 10 preguntas. Posteriormente se estima un punto de corte por cada panelista, y finalmente se calcula el promedio de los puntos de corte estimados para fijar cuál sería este.

Tabla 2: Ejemplo de método Angoff

Pregunta	Panelista 1	Panelista 2	Panelista 3	Panelista 4
1	0.90	0.85	0.75	0.80
2	0.65	0.65	0.60	0.55
3	0.85	0.90	0.85	0.80
4	0.65	0.60	0.65	0.60
5	0.55	0.50	0.45	0.45
6	0.60	0.80	0.85	0.90
7	0.75	0.70	0.70	0.85
8	0.80	0.90	0.75	0.90
9	0.65	0.55	0.50	0.50
10	0.80	0.95	0.90	0.95
Punto de corte por panelista	7.20	7.40	7.00	7.30
Promedio de punto de corte			7.23	

Para 1990, el método Angoff continuaba como el método de elección para establecer un punto de corte, era ampliamente utilizado y estudiado; sin embargo, fue atacado debido a que expertos consideraban que este y todos los métodos basados en la evaluación de las preguntas eran poco factibles de desarrollar. Algunos de los argumentos más fuertes en contra de estos métodos son: la alta inversión de costo y tiempo, además de esto, el ensamblar un panel de expertos no es nada fácil, ya que se debe asegurar una representatividad que impactará la definición del estándar. Una vez armado, algunos de los retos son asegurar que los panelistas sean capaces de hacer lo que este método requiere, y estimar la probabilidad de que un estudiante hipotético conteste correctamente (Katz y Tannenbaum, 2014). No obstante, es común que el estándar haya sido fijado muy alto, por lo que se han propuesto distintas modificaciones en la que los panelistas tienen permitido ajustar sus estimados después de la *dosis de realidad* de ver las respuestas de los participantes (Schoonheim-Klein et al. 2009).

Por otro lado, diversos estudios han reportado implementaciones donde los jueces involucrados en las revisiones de diferentes evaluaciones se sentían confiados con su habilidad para realizar las tareas del método Angoff, entendían que tenían que hacer y cómo hacerlo y creían que los resultados eran defendibles y suficientemente buenos (Papageorgiou y Tannenbaum, 2016). Este proceso para establecer un estándar es utilizado en la actualidad en diferentes pruebas de lenguaje como lo es el examen de inglés como segundo idioma (TOEFL), que es una prueba estandarizada de dominio del idioma inglés. Algunas innovaciones que han surgido se enmarcan en la inclusión de tecnología para llevar a cabo las sesiones con panelistas en la virtualidad y hacer el análisis de forma más eficiente (Katz y Tannenbaum, 2014).

MÉTODO NEDELSKY

En 1954, Leo Nedelsky sugirió un método de evaluación en el cual se determinan estándares de calificación absolutos para pruebas objetivas, específicamente aquellas de opción múltiple. Este método se volvió muy popular no solo en exámenes escolares sino también en certificaciones profesionales y licenciaturas (Chang, 2009). Nedelsky se basaba en la idea de conceptualizar a los estudiantes en el límite entre aprobar y reprobar (llamándolos los estudiantes F-D), e identificar las opciones de las preguntas de opción múltiple que estos estudiantes serían capaces de eliminar como incorrectas. El número de opciones remanentes es la probabilidad de que el estudiante mínimamente competente conteste correctamente. La suma de las probabilidades se considera el puntaje esperado para este estudiante y es la base para el punto de corte.

El método de Nedelsky es popular en contextos médicos, presuntamente porque el evaluado, debe de ser capaz de rechazar esas opciones que "causarían daño a un paciente". A lo largo del método de Nedelsky existen dos suposiciones que podrían influir en los resultados, las cuales son, que los jueces sean capaces de discriminar entre las opciones que los evaluados pudieran considerar como opción y aquellas que no, y que los evaluados pueden responder correctamente contestando de manera aleatoria (Zieky, 2001).

Un ejemplo del método de Nedelsky se muestra en la Tabla 3, la cual muestra una evaluación de una prueba con 10 preguntas de opción múltiple, en ella el juez descarta las opciones prediciendo que una persona mínimamente competente sería capaz de identificar esas respuestas como incorrectas. Por lo tanto, en la segunda pregunta la probabilidad sería de 0.50, la recomendación de este juez sobre el punto de corte es de 4.83.

Tabla 3: Ejemplo de método Nedelsky

Pregunta	Opciones					Restantes	Probabilidad
1	~~A~~	~~B~~	C	~~D~~	~~E~~	1	1.00
2	A	B	~~C~~	~~D~~	~~E~~	2	0.50
3	~~A~~	~~B~~	C	D	E	3	0.33
4	A	B	C	~~D~~	E	4	0.25
5	A	B	C	D	~~E~~	4	0.25
6	A	~~B~~	~~C~~	D	E	3	0.33
7	~~A~~	~~B~~	~~C~~	~~D~~	E	1	1.00
8	A	~~B~~	~~C~~	~~D~~	E	2	0.50
9	~~A~~	~~B~~	C	D	E	3	0.33
10	A	B	~~C~~	~~D~~	E	3	0.33
Recomendación que este juez hace sobre el punto de corte							4.83

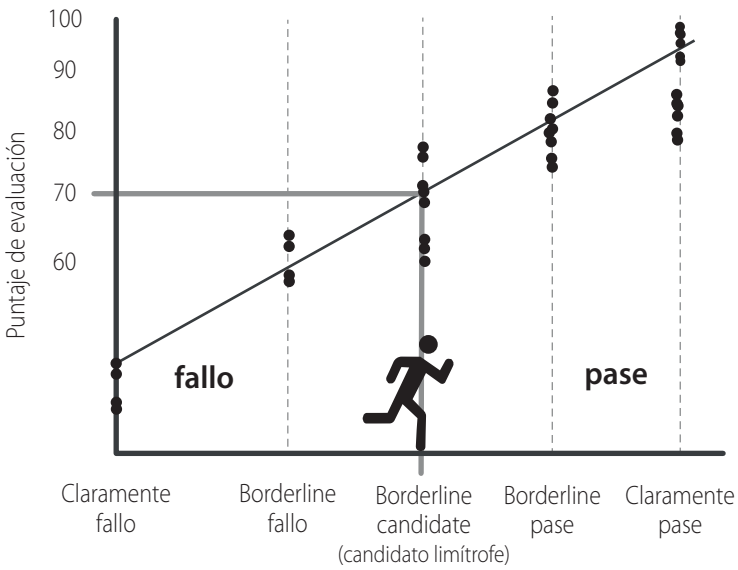
BORDERLINE GROUP

El método Borderline Group es un método centrado en el examinado en lugar de un método centrado en la prueba, por lo que los juicios se hacen sobre los examinados individuales, no sobre el contenido del examen. El método se puede utilizar solo cuando los expertos en contenido que están calificados para servir como emisores de estándares, como lo son los profesores o expertos, observan de manera directa la prueba de desempeño. Las calificaciones globales de los jueces observadores se utilizan para determinar la puntuación de la lista de verificación que se utilizará como estándar de aprobación.

Este método, como el de Nedelsky y Angoff, se basa en la premisa de que la calificación aprobatoria debe ser aquella obtenida por cualquier evaluado mínimamente competente; sin embargo, en lugar de evaluar cada una de las preguntas u opciones, los jueces evalúan directamente a quienes tomen la prueba y posteriormente, la calificación media se toma como calificación aprobatoria (Schoonheim-Klein, 2009).

El procedimiento para llevar a cabo este método inicia mediante la preparación de los jueces orientándolos en el caso y los instrumentos para calificar. Posteriormente los jueces observan directamente el rendimiento de todos los evaluados. Como se muestra en la Figura 2, los jueces proveen una calificación global del rendimiento general de cada uno en una escala de tres puntos, no aprobado (fallo), borderline (limítrofe) y aprobado (pase).

Figura 2: Ejemplo del Método Borderline



Una debilidad del método es que el número de examinados clasificados como dentro del límite es a menudo pequeño. En la mayoría de las situaciones, si es posible recopilar calificaciones, también será posible recopilar datos para implementar el método de grupos contras-

tantes. Si es así, es preferible contrastar grupos porque los datos están directamente relacionados con la decisión a tomar.

Implicaciones, problemas y críticas hacia los modelos existentes

Sin importar el método o modelo elegido para la definición del estándar, este involucra juicios subjetivos de expertos, análisis psicométrico, definición de políticas para la toma de decisiones y la socialización de estos resultados (Papageorgiou y Tannenbaum, 2016). Es por eso por lo que es una tarea inseparable, el definir criterios claros, justos y defendibles para que deje de ser una tarea tan controvertida (Downing et al., 2010).

Una crítica para la definición de estándares es que estos fueron fijados en ciertas condiciones históricas, sociales y culturales, asumiendo un ideal que difícilmente es alcanzable en las pruebas de examinación donde los participantes tienen mucho en juego, por ejemplo, en aquellas de ubicación en un programa de idiomas o aquellas que garantizan la entrada a una plaza de estudios de posgrado.

Otro problema altamente relacionado se refiere a la interpretación de resultados de los diferentes grupos de interés. Frecuentemente, los resultados se usan en dos niveles de interpretación, el primero alrededor de la estimación de las habilidades de los participantes en la evaluación, y el segundo para fines de evaluación de la calidad de los programas formativos. Si el foco es este, los estándares no deben ser la excusa para señalar culpables o fincar responsabilidades a estudiantes, escuelas o países, ya que los datos deben tener un análisis agregado y la reflexión enfocarse en los contextos donde estas evaluaciones están teniendo lugar.

IMPLICACIONES DE LA PANDEMIA Y EVALUACIÓN A DISTANCIA

La pandemia del COVID-19 ha traído consigo muchos retos para la educación, desde la adaptación de un programa educacional a un ambiente virtual de enseñanza, el mantenernos cerca a los estudiantes, así como el diseño de actividades y evaluaciones.

Los grandes cambios que hemos experimentado al traer la escuela o la oficina a la casa, se han traducido en la creación de ambientes compartidos donde los alumnos enfrentan la dificultad para ajustarse a un nuevo método de estudio, responsabilidades en el hogar, y distracciones como ruidos u otro tipo de interrupciones que afectan su concentración.

Aunque hablar de educación híbrida o a distancia se ha convertido en algo común en los diferentes niveles educativos, como el hecho que formará parte de una nueva normalidad incluso una vez que el virus no represente una amenaza para la salud; la calidad y la validez de una educación remota continúa siendo un tema controversial.

Particularmente, la discusión gira alrededor del constante reto de la educación, la evaluación. En primer lugar, cómo distribuirla y llevarla a cabo de manera exitosa, además de cómo asegurar la integridad o disminuir las oportunidades para deshonestidad escolar.

McCabe et al. (2012), señalan que la incidencia de deshonestidad académica ha ido en aumento en los últimos años, adjudicándose a la facilidad de acceso a fuentes electrónicas

que tienen las nuevas generaciones. Debido a esto, distintas autoridades escolares han tomado diferentes acciones para combatir la deshonestidad en pruebas académicas incluyendo el uso de la tecnología. En particular, se ha hecho popular el uso de programas y plataformas que evitan que el alumno pueda acceder a otras ventanas o bloquean el uso de internet después de haber descargado el examen. Esto resuelve los problemas para exámenes de opción múltiple enfocados en la memorización de información, pero fallan en considerar otros factores para valorar el desempeño integral del alumno.

Junto a esto se han utilizado técnicas como mantener el micrófono y cámara encendidos en todo momento, el uso de varios dispositivos para que el evaluador pueda visualizar todo el ambiente en el cual se realiza el examen e incluso la reducción del tiempo en el examen para evitar que se busque información por otros medios. Sin embargo, aunque estas técnicas pueden parecer atractivas, dejan de lado las dificultades del acceso a la tecnología que muchos de nuestros estudiantes han enfrentado. Junto a la labor de evitar que los alumnos incurran en una deshonestidad, también es importante considerar el acceso a recursos de aprendizaje, así como internet y equipos de cómputo, y los básicos como agua o electricidad, que pueden impactar en la valoración del desempeño que se hace del alumno. Definir estándares de desempeño debe ser sensible también a los contextos alrededor del diseño, implementación y análisis de las pruebas.

CONCLUSIONES Y REFLEXIONES FINALES

La definición de estándares de evaluación está ligada a los currículos basados en competencias, donde las universidades y los programas educativos deben garantizar un nivel mínimo de aptitud. Los métodos discutidos previamente fueron establecidos hace más de cincuenta años, con el fin de encontrar una forma de evaluar de manera estándar a un grupo ampliado de estudiantes. En la actualidad, el foco de la educación ha cambiado, ya no es la educación masificada, ahora el enfoque es una educación flexible y personalizada, los retos que enfrentamos han cambiado y el modo de evaluar debe evolucionar también, tomando en cuenta diferentes factores para valorar un buen desempeño.

A lo largo de la pandemia del COVID-19, los retos cambiaron y las evaluaciones de manera remota han demostrado que pueden tener altas repercusiones aun en el rendimiento de los alumnos. Estas posibles repercusiones incluyen la presión añadida de alcanzar una nota alta, por lo que algunos alumnos han sentido también la necesidad de cometer deshonestidades académicas. A pesar de que la tecnología ha facilitado el aprendizaje y ha sido la base de la educación remota, la evaluación sigue siendo un reto debido a la imposibilidad de aplicar estándares a todos los alumnos sin considerar el ambiente en casa o las disparidades en el acceso a la educación a distancia.

REFERENCIAS

Barman, A. (2008). Standard setting in student assessment: is a defensible method yet to come? *Annals Academy of Medicine Singapore, 37*(11), 958-963.

Boud, D. (2006). Foreword. En C. Bryan & K. Clegg (Eds.), Innovative assessment in higher education (pp. XVII–XIX). London and New York: Routledge.

Cetin, S., Gelbal, S. (2013). A Comparison of Bookmark and Angoff Standard Setting Methods. *Educational Sciences: Theory and Practice, 13*(4), 2169-2175.

Chang, L. (2009). Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods. *Applied Measurement in Education, 12*(2), 151-165.

Cusimano, M.D. (1996). Standard Setting in Medical Education. *Academic Medicine, 71*(10), 112-121.

Downing, S. M., Tekian, A., Yudkowksy, R. (2010). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine, 18*(1), 50-57.

Joughin, G. (2008). Assessment, Learning and Judgement in Higher Education: A Critical Review. En: G. Joughin (eds) Assessment, Learning and Judgement in Higher Education, 1–15.

Kampa, N., Wagner, H., Köller, O. (2019). The standard setting process: validating interpretations of stakeholders. *Large-scale Assessments in Education, 7*, 3. https://doi.org/10.1186/s40536-019-0071-8

Katz, I., Tannenbaum, R.J. (2014). Comparison of Web-based and Face-to-face Standard Setting using the Angoff method. *Journal of Testing Technology, 15*(1),1-17.

Lázaro, J.L., Usart, M., Gisbert, M. (2019). La evaluación de la competencia digital docente: construcn ción de un instrumento para medir los conocimientos de futuros docentes. Journal of New Approaches in Educational Research, *8*(1), 75-81.

McCabe, D. L., Butterfield, K. D., & Trevino, L. K. (2012). *Cheating in college: Why students do it and what educators can do about it*. JHU Press.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating Standard Setting within Argument-Based Validity. Language Assessment Quarterly, *13*(2), 109–123. https://doi.org/10.1080/15434303.2016.1149857

Ruano, A.L., Vizuete, A., Moreno, J.C., Quispe, W. (2018). Habilitación profesional: caso Ecuador. Rev. Ecu. Med. Eugenio Espejo. *7*(9), 15-20.

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., Van Der Vleuten, C., & Van Der Velden, U. (2009). Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. European Journal of Dental Education, *13*(3), 162–171. https://doi.org/10.1111/j.1600-0579.2008.00568.x

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. En G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 19–51). Lawrence Erlbaum Associates Publishers.