MINISTERIO DE EDUCACIÓN PÚBLICA

GOBIERNO DE COSTA RICA

DGEC
Dirección de Gestión y Evaluación de la Calidad

Marco de referencia

Pruebas Nacionales Estandarizadas Diagnósticas 2026

Primaria y secundaria

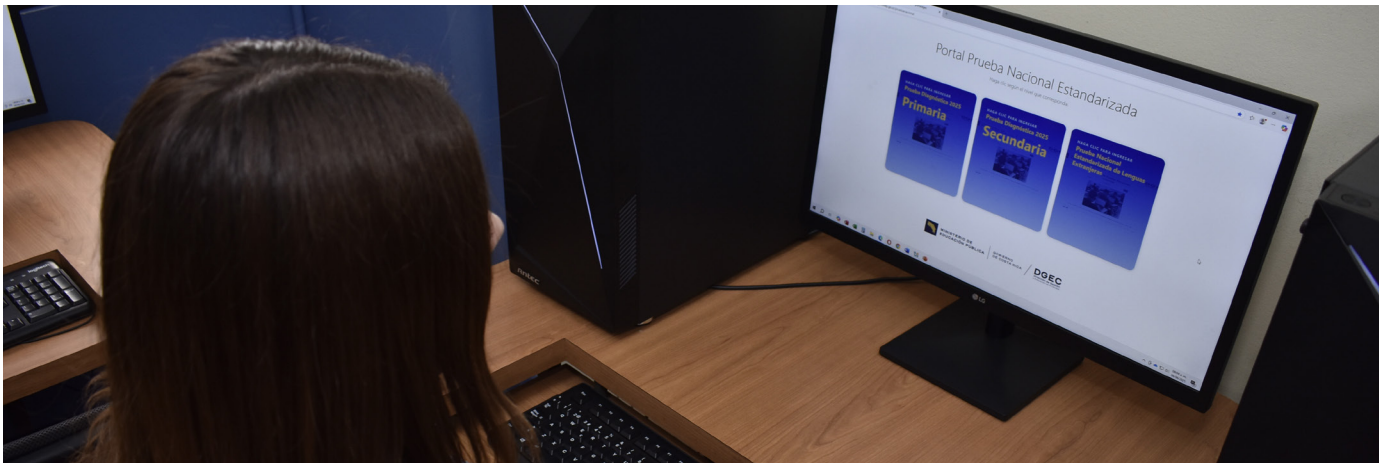
MINISTERIO DE EDUCACIÓN PÚBLICA | GOBIERNO DE COSTA RICA | DGEC Dirección de Gestión y Evaluación de la Calidad

Trabajo elaborado por colaboradores de la Dirección de Gestión y Evaluación de la Calidad:

Coordinación técnica: Álvaro Artavia Medrano
Maquetación y diseño: Johan Herra González

Tabla de contenidos

Presentación



El Decreto Ejecutivo No. 45509-MEP, en su título II, presenta la normativa propia de las pruebas nacionales del sistema educativo costarricense.

Específicamente, en el artículo 83, se establece que las pruebas nacionales constituyen un conjunto de instrumentos de evaluación administrados por la Dirección de Gestión y Evaluación de la Calidad (DGEC) del Ministerio de Educación Pública (MEP), según los propósitos definidos por el Consejo Superior de Educación.

En particular, las **Pruebas Nacionales Estandarizadas por asignatura, con propósito diagnóstico en primaria y secundaria para el curso lectivo 2026**, tienen la finalidad de determinar los niveles de logro de los aprendizajes por parte de las personas estudiantes.

En el ámbito de las evaluaciones de gran escala, el **Marco de Referencia** constituye una garantía de transparencia, coherencia y rigor técnico. Al delimitar con precisión el objeto de evaluación, su fundamento curricular y el modelo de medición adoptado, explicita con claridad qué se evalúa, por qué se evalúa y bajo qué criterios se interpretan los resultados. De esta manera, asegura que las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico respondan a principios de validez, equidad y consistencia, y que la información generada se utilice de forma responsable para orientar la mejora continua del sistema educativo.

En este documento, se presenta el marco conceptual y técnico que orienta de manera integral las **Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico para el curso lectivo 2026**. No se trata únicamente de un documento descriptivo, sino del punto de partida que da sentido y coherencia a los procesos relacionados con estas pruebas, asegurando una lectura informada y contextualizada de la macroevaluación en el sistema educativo nacional.

Marco de **referencia**

El **marco de referencia** de las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico para el curso lectivo 2026 establece los fundamentos conceptuales, curriculares y técnicos que delimitan el alcance de la evaluación. En ese sentido, brinda una noción clara del objeto de evaluación, su extensión y complejidad. En particular, es un requisito técnico y de transparencia en el diseño e implementación de evaluaciones válidas y confiables (Rodríguez Frías & Flotts de los Hoyos, 2019).

En el contexto de las Pruebas Nacionales Estandarizadas por asignatura, el **marco de referencia** articula de manera sistemática la Política Educativa, la Política Curricular, los programas de estudio oficiales y el modelo de medición y de evaluación adoptados. Esta articulación asegura que la macroevaluación responda estrictamente al currículo vigente y que la estimación del desempeño se sustente en aprendizajes formalmente establecidos. De esta manera, el marco de referencia evita ambigüedades, previene interpretaciones discrecionales y delimita con claridad los alcances de la prueba.

Algunos aspectos que reflejan la utilidad del marco de referencia para las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico son:

- Orientar el proceso evaluativo.
- Definir y delimitar con claridad el objeto de evaluación.
- Garantizar coherencia en los distintos procesos implicados en la evaluación.
- Establecer el modelo de medición y evaluación que permitirá generar la prueba, así como el correcto análisis e interpretación de los resultados.

Desde el punto de vista estructural y teórico, el **marco de referencia** organiza los elementos que permiten pasar del currículo a la medición. Incluye, entre otros aspectos, el análisis del dominio, el establecimiento de bloques temáticos, la definición de afirmaciones, la determinación de evidencias, el desarrollo de tareas, el modelo de evaluación y el modelo de medición.

Esta secuencia no es solamente formal, sino que constituye el argumento técnico que vincula lo que el currículo prescribe con lo que efectivamente se observa en las respuestas del estudiantado a los ítems de las pruebas nacionales.

Propósito y objetivos de las **Pruebas Nacionales Estandarizadas**

De acuerdo con el artículo 84 del Decreto Ejecutivo No. 45509-MEP "Reglamento de Evaluación de los Aprendizajes y la Conducta" (REAC), las **Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico** permiten evidenciar el logro de aprendizajes por parte de las personas estudiantes, con el fin de generar insumos para la mejora del proceso de aprendizaje y enseñanza en particular y de la calidad del sistema educativo, en general.

Las Pruebas Nacionales Estandarizadas son elaboradas y aplicadas por asignatura y tienen como propósito monitorear el desarrollo de aprendizajes esenciales de los estudiantes para la mejora continua en el proceso de enseñanza y aprendizaje, entre el inicio y el final del año académico, asimismo, son requisito para la obtención del Certificado de Conclusión de Estudios de Primero y Segundo Ciclos de la Educación General Básica, y el Título de Bachiller en Educación Media.

En su propósito diagnóstico, ese mismo artículo indica que las pruebas tienen como finalidad la determinación de los niveles de logro de los aprendizajes por parte de las personas estudiantes. Asimismo, su aplicación es censal y obligatoria, durante el primer periodo o semestre del curso lectivo, según corresponda a la modalidad u oferta educativa.

Finalmente, en el artículo 88 del Decreto No. 45509-MEP se establecen los objetivos de las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico:

a) Contribuir al proceso de mediación pedagógica para la adquisición de habilidades, conocimientos, actitudes y valores en las personas estudiantes.

b) Coadyuvar en el proceso de enseñanza y aprendizaje orientado al logro de las competencias requeridas en la formación ciudadana:

   - Competencias para la vida en ciudadanía
   - Competencias para la vida: sociales, emocionales y de aprendizaje.
   - Competencias para el empleo digno y el emprendimiento.

c) Generar los insumos a partir de los resultados de la prueba para el seguimiento de estrategias de mejora continua del proceso de enseñanza y aprendizaje, de la mediación pedagógica y la evaluación de los aprendizajes.

Las **Pruebas Nacionales Estandarizadas por asignatura** con propósito diagnóstico evaluarán conocimientos, habilidades y otras destrezas propias de las asignaturas según los enfoques y fundamentación teórica de los programas de estudio de las asignaturas de Español, Matemáticas, Ciencias (Biología, Física y Química), Estudios Sociales y Educación Cívica (secundaria).

Formatos de aplicación

Las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico tendrán dos formatos de aplicación: digital y físico, de acuerdo con el artículo 98 del REAC.

La DGEC emitirá los criterios técnicos y administrativos correspondientes a cada formato de aplicación. En lo que respecta a la **aplicación digital**, esta se llevará a cabo en coordinación con las direcciones regionales de educación, considerando la infraestructura y el equipamiento tecnológico requeridos para garantizar la gestión de la prueba en línea, su aplicación, sin excepción, en ambientes controlados y supervisados en las sedes designadas, así como la implementación de los instructivos y protocolos emitidos por la DGEC para tal fin.

En cuanto al **formato físico**, se toman en consideración aspectos tales como no contar con el equipamiento tecnológico requerido, incidencias en el servicio eléctrico, centros educativos que imparten ofertas para personas privadas de libertad, Centro de Apoyo en Pedagogía Hospitalaria, o bien, situaciones de caso fortuito o de fuerza mayor que deberán ser valoradas por la Dirección de Gestión y Evaluación de la Calidad.

Duración de las pruebas

De acuerdo con el artículo 94 del REAC, las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico no sobrepasarán los 180 minutos por asignatura. **Para el curso lectivo 2026, la duración será de 120 minutos en cada prueba**. En el caso de estudiantes que posean el apoyo educativo de una hora adicional, la población estudiantil que así lo requiera contará con 180 minutos.

Modelo de **evaluación**

El **diseño centrado en evidencias** considera la evaluación como un argumento probatorio, esto es, un argumento que parte de lo que se observa que el estudiantado dice, hace o realiza en unas circunstancias concretas, para inferir lo que saben, pueden hacer o han logrado de forma más general (Mislevy et al., 2003).

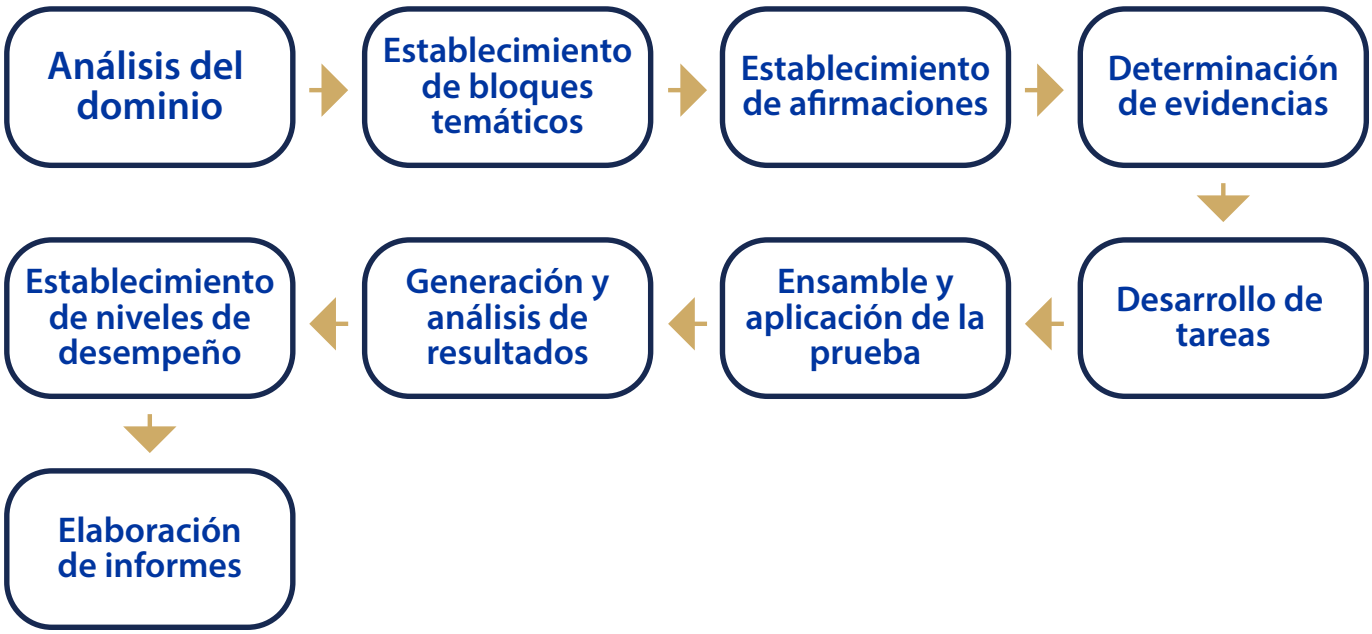
El diseño centrado en evidencias es un marco estructural para analizar y desarrollar evaluaciones como un ejercicio de razonamiento probatorio, con la validez como base para las inferencias que se pretendan hacer a partir de la información recolectada (Mislevy et al, 2017).

Este modelo se considera un enfoque lógico y sistemático que organiza el trabajo de diseño e implementación de la evaluación en términos de capas o niveles; una metáfora extraída de la arquitectura y la ingeniería de software (Mislevy y Riconscente, 2005). Cada uno de estos niveles cuenta con conceptos clave y un propósito definido que, aunque sugieren un proceso de diseño secuencial, en realidad pueden interactuar, actualizarse e ir mejorando.

En la figura 1 se muestran los pasos vinculados con los niveles del diseño centrado en evidencias, los cuales se pueden interpretar como una cadena de razonamiento en cuanto a argumentaciones y garantías acerca de las inferencias que se puedan hacer sobre los resultados de la población estudiantil en la prueba.

Figura 1

Pasos por seguir de acuerdo con el diseño centrado en evidencias y otros procesos técnicos relacionados con el ensamblaje de las Pruebas Nacionales Estandarizadas por asignatura 2026



A continuación, presentamos la descripción de los niveles adoptados para el diseño, a partir de Mislevy et al. (2017), Mislevy et al. (2003) y Zieky (2014) y otros procesos relacionados con la implementación de las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico 2026.

Análisis del dominio

Las **Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico** para el curso lectivo 2026 son elaboradas y aplicadas por asignatura, de acuerdo con el artículo 84 del REAC. Evaluarán conocimientos y habilidades propias de las asignaturas objeto de medición. El principal insumo para el diseño y construcción de las pruebas serán los enfoques y fundamentación teórica de los programas de estudio vigentes.

Para tales propósitos, en el nivel **"Análisis del dominio"** el equipo técnico-asesor de cada asignatura realizó una revisión curricular de la Política Educativa y de la Política Curricular vigentes, así como de la fundamentación de los programas de estudio y los conocimientos y las habilidades que son medibles en una prueba constituida únicamente por ítems de selección única. De esta manera, es posible especificar los conceptos, la terminología, las formas de representación y las maneras de interactuar en relación con los conocimientos y las habilidades que son objeto de evaluación, luego de haber comprendido su relevancia en la formación del estudiantado, la forma en que son representados en el currículo, la relación entre ellos y la manera en que son adquiridos como parte del proceso educativo formal.

De acuerdo con Mislevy y Riconscente (2005), las investigaciones sobre el aprendizaje también nos dicen mucho sobre cómo las personas llegan a ser competentes en determinados ámbitos y, por tanto, sobre lo que debemos evaluar.

Como primer paso en el diseño de la evaluación, el análisis del dominio nos permite comprender los conocimientos y las habilidades que las personas estudiantes utilizan en una asignatura específica, así como las características de las situaciones que evoquen el uso de conocimientos disciplinares, procedimientos y estrategias.

Establecimiento de los bloques temáticos

Un **bloque temático** se refiere a una agrupación de habilidades y conocimientos que están interrelacionados con respecto a una temática en particular en cada asignatura objeto de medición en las pruebas.

Se presentan de forma organizada, coherente, con una secuencia lógica y pedagógica, producto de un análisis curricular previo de los programas de estudio vigentes, así como de la Política Educativa y la Política Curricular.

Los **bloques temáticos** toman en cuenta factores como la progresión del aprendizaje y la relevancia de las temáticas. Además, guardan congruencia con los demás niveles del diseño centrado en evidencias, asumido como modelo de evaluación para estas pruebas.

Establecimiento de afirmaciones

De acuerdo con Zieky (2014), el **diseño basado en evidencias** se puede considerar como un medio para construir una cadena de argumentos y garantías que respaldan las afirmaciones que se hacen sobre las personas estudiantes, a partir de una prueba.

Las **afirmaciones** son enunciados acerca de lo que las personas estudiantes muestran en su desempeño en la prueba en cada asignatura, por lo que deben ser claras y precisas, así como deben apoyarse en otras afirmaciones que tengan diferentes niveles de especificidad.

El propósito de una prueba y las **afirmaciones** que se pueden hacer sobre el estudiantado están estrechamente relacionados. En otras palabras, el propósito de la prueba se puede expresar como la afirmación de mayor nivel que se puede formular a partir de los resultados.

Determinación de evidencias

Las **evidencias** se basan en los comportamientos o productos observables que se pueden concretar en las respuestas a una tarea específica. Para ello, se debe hacer una descripción detallada de cuáles son los conocimientos y las habilidades en los que se centra la medición para cada una de las asignaturas.

En este modelo de evaluación, las **evidencias** brindan las garantías que permiten apoyar o fundamentar las afirmaciones que se han establecido ya que se enfocan en aspectos relevantes de los comportamientos o productos observables, así como la diferenciación de lo que es medible en una prueba con respecto a lo que no lo es.

Para determinar evidencias se puede partir de una situación ideal o hipotética, no obstante, las restricciones o condiciones reales de cada prueba (por ejemplo, formato de aplicación, tiempo reglamentario, entre otros), así como la relevancia para respaldar una afirmación son aspectos cruciales para emprender el siguiente paso.

Desarrollo de tareas

El **desarrollo de tareas** se basa en una descripción de lo que cada prueba incluirá para la medición específica del objeto de evaluación por asignatura. En particular, se enfocará en el tipo de ítems que son situaciones que provocan el comportamiento o producto observable tal y como fue descrito en las evidencias (Zieky, 2014).

En la sección correspondiente se detallará el tipo de ítems que conforman las **Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico**, así como otras características referentes a la cantidad de opciones.

Enfoque para la **interpretación de resultados**

La interpretación de resultados de la aplicación diagnóstica de las Pruebas Nacionales Estandarizadas por asignatura se realiza desde un **enfoque de evaluación referida a criterios**, lo que implica que el desempeño del estudiantado se interpreta en función de aprendizajes previamente definidos en el currículo oficial y no en comparación con el rendimiento de otros estudiantes. En este modelo, los resultados se valoran con base en el grado de dominio alcanzado respecto de afirmaciones y evidencias claramente establecidas en el presente marco de referencia, garantizando coherencia entre lo que se enseña, lo que se espera que se aprenda y lo que efectivamente se evalúa.

Este enfoque permite describir con precisión qué conocimientos, habilidades y procesos cognitivos han sido consolidados y cuáles requieren fortalecimiento, favoreciendo una lectura pedagógica de los resultados. En otras palabras, este enfoque consiste en "privilegiar la comparación del desempeño de un individuo con una definición clara y precisa de lo que se espera que conozca y sea capaz de hacer en un determinado dominio" (Ravela, 2006, p. 43).

La **evaluación referida a criterios** permite describir con claridad qué aprendizajes han sido consolidados y cuáles requieren fortalecimiento, favoreciendo una lectura pedagógica de los resultados. Al no depender de comparaciones con el rendimiento promedio de otros estudiantes ni de posiciones relativas dentro de un grupo, la evaluación referida a criterios ofrece una interpretación más justa y centrada en el grado de dominio alcanzado. De esta manera, los resultados se comprenden en función de los aprendizajes esperados y no en función de quién obtiene mejores o peores puntuaciones dentro del conjunto evaluado.

Definición del objeto de **evaluación**

El **objeto de evaluación** de las Pruebas Nacionales Estandarizadas por asignatura está constituido por el desempeño observable del estudiantado en tareas que requieren la activación integrada de conocimientos disciplinares, habilidades y procesos cognitivos vinculados a las competencias curriculares establecidas en los programas de estudio vigentes.

Las pruebas diagnósticas por asignatura no miden contenidos aislados, sino la capacidad de aplicar conocimientos en contextos estructurados que demandan:

- Pensamiento sistémico
- Pensamiento crítico
- Resolución de problemas

Estas habilidades actúan como dimensiones cognitivas transversales que se operacionalizan dentro de cada campo disciplinar.

En consonancia con los planteamientos de la Política Curricular acerca de una evaluación transformadora que "se asuma como una forma de identificar la complejidad de los retos y los nuevos elementos que se integran a los nuevos aprendizajes" (MEP, 2015, p. 27), se cuenta con pruebas que corresponden a los conocimientos disciplinares vinculados a las competencias curriculares.

En cada asignatura evaluada, la prueba se fundamenta en la definición explícita de su **objeto de evaluación** o **constructo**, entendido como el conjunto estructurado de conocimientos, habilidades y procesos cognitivos que se pretende medir de manera integrada. La delimitación del constructo permite precisar qué dimensiones del aprendizaje serán consideradas, establecer los límites conceptuales de la medición y garantizar coherencia entre el currículo oficial, el marco de especificaciones y el modelo de medición adoptado. De esta manera, el **objeto de evaluación** constituye el referente conceptual que orienta la construcción de los ítems, la organización de la prueba y la interpretación diagnóstica de los resultados.

Primaria

Ciencias

Es la capacidad de comprender fenómenos del entorno mediante la interpretación, el análisis y el contraste de información científica, y de elaborar y sostener explicaciones o conclusiones basadas en evidencia. También incluye la capacidad de tomar decisiones fundamentadas ante situaciones del entorno, de manera que se anticipen consecuencias a partir de la evidencia disponible, en distintos contextos.

Español

Se concreta en la competencia de analizar e integrar información de textos escritos (literarios y no literarios) mediante la aplicación de estrategias cognitivas que permitan construir significado literal, inferencial y crítico. Este dominio se evidencia no solo en la capacidad de acceder al significado literal de los textos, sino en la comprensión subyacente mediante estrategias que permiten integrar información relevante, distinguir ideas fundamentales de las complementarias, determinar las causas y los efectos, reconocer los temas tratados, conflictos y comportamientos, así como inferir los pensamientos de los personajes; todo esto acorde con las demandas comunicativas y según los contextos respectivos.

Estudios Sociales

Se concibe como la habilidad para integrar y movilizar conocimiento social, articulando de forma conjunta las perspectivas histórica, geográfica y ciudadana, a fin de interpretar situaciones del entorno, explicar relaciones entre sociedad, espacio y tiempo, y tomar y justificar decisiones mediante pensamiento crítico, orientadas a la convivencia democrática y la sostenibilidad.

Matemáticas

Es la capacidad para formular, emplear e interpretar las Matemáticas en una variedad de contextos. Incluye razonar matemáticamente y usar conceptos, procedimientos, hechos y herramientas para describir, explicar y predecir fenómenos. Ayuda a reconocer el papel de las Matemáticas en el mundo y hacer juicios bien fundados y decisiones necesarias para ciudadanos constructivos, comprometidos y reflexivos.

Secundaria

Ciencias
Es la capacidad para utilizar conocimientos y prácticas de las Ciencias con el fin de comprender y explicar fenómenos y situaciones, interpretar información y evidencias, analizar representaciones simbólicas y gráficas, así como sustentar conclusiones o decisiones fundamentadas relacionadas con la comprensión de los sistemas vivos, sus procesos, interacciones y el ambiente; la interpretación de la estructura, propiedades y transformaciones de la materia; y el análisis de los principios que explican el movimiento, la energía, las fuerzas y otras interacciones que rigen los fenómenos físicos.

Educación Cívica
Se concibe como la capacidad para analizar críticamente los regímenes políticos, comprender la organización y funcionamiento del sistema democrático costarricense, interpretar el sistema electoral y las políticas públicas desde la perspectiva de la igualdad de oportunidades, y movilizar valores, actitudes y prácticas democráticas para ejercer una ciudadanía activa, responsable y participativa en distintos contextos sociales.

Español
Como constructo teórico, la comprensión de lectura y análisis literario se concibe como una competencia cognitiva que permite a las personas estudiantes construir sentido a partir de textos literarios y no literarios, reorganizar la comprensión textual, interpretar la posición del texto, interpretar recursos retóricos, inferir las implicaciones del pensamiento social e interpretar significados explícitos e implícitos para responder a propósitos lectores diversos. Acá se focaliza la capacidad para comprender, interpretar, evaluar y analizar textos literarios y no literarios, con el fin de alcanzar objetivos personales, desarrollar conocimientos, participar activamente en la sociedad y responder eficazmente a demandas académicas y sociales.

Estudios Sociales
Se concibe como la habilidad para integrar y aplicar conocimientos sociales de forma crítica y reflexiva al articular las perspectivas históricas, geográficas, económicas, políticas, ambientales y culturales. Esta habilidad se manifiesta cuando el estudiantado localiza y contrasta información, establece relaciones de causa y consecuencia y de multicausalidad, compara procesos y actores sociales en diferentes tiempos y espacios y analiza problemas contemporáneos a partir de diversas fuentes. Mediante estas tareas, se evidencia una comprensión fundamentada relacionada con la convivencia democrática, el respeto de los derechos humanos, la diversidad sociocultural y la sostenibilidad, en coherencia con el currículo nacional.

Matemáticas
Es la capacidad para formular, emplear e interpretar las Matemáticas en una variedad de contextos. Incluye razonar matemáticamente y usar conceptos, procedimientos, hechos y herramientas para describir, explicar y predecir fenómenos. Ayuda a reconocer el papel de las Matemáticas en el mundo y hacer juicios bien fundados y decisiones necesarias para ciudadanos constructivos, comprometidos y reflexivos.

Modelo de medición

El modelo de medición adoptado en las Pruebas Nacionales Estandarizadas sustenta técnicamente la estimación e interpretación del desempeño del estudiantado, garantizando precisión, coherencia y validez en los resultados. En el marco de la aplicación diagnóstica, este modelo integra aportes de la **Teoría Clásica de los Test (TCT) y de la Teoría de Respuesta a los Ítems (TRI)**, enfoques complementarios que permiten analizar la calidad de los ítems, estimar la consistencia de la prueba y ubicar el desempeño en una escala continua de habilidad. La articulación de ambos enfoques fortalece el carácter referido a criterios de la evaluación y respalda la conformación técnica de los niveles de desempeño.

En este marco, la TCT se utiliza con fines diagnósticos y descriptivos para examinar la calidad técnica inicial de ítems y pruebas, mediante indicadores como dificultad observada, discriminación, consistencia interna y funcionamiento de distractores. El modelo de Rasch constituye el fundamento principal para la calibración de ítems, la estimación de habilidad, el análisis del error estándar de medición, el establecimiento de niveles de desempeño y la construcción del indicador de reporte.

Teoría Clásica de los Test

En el contexto de la aplicación diagnóstica de las Pruebas Nacionales Estandarizadas por asignatura, los aportes de la **Teoría Clásica de los Test** permiten realizar análisis detallados del comportamiento de los ítems y del desempeño global del estudiantado. A partir de este enfoque se examinan indicadores como el índice de dificultad, que describe la proporción de estudiantes que responde correctamente cada ítem; el índice de discriminación, que refleja la capacidad del ítem para diferenciar entre distintos niveles de dominio; y la medida de la consistencia interna como evidencia de la confiabilidad de las puntuaciones obtenidas. Estos análisis resultan fundamentales para asegurar que la prueba proporcione información estable y técnicamente sólida.

La aplicación diagnóstica, al tener como finalidad orientar la mejora pedagógica, requiere que los ítems no solo sean válidos en términos de contenido, sino que además funcionen adecuadamente desde el punto de vista psicométrico. El análisis clásico permite identificar ítems demasiado fáciles o excesivamente difíciles, detectar distractores que no cumplen una función adecuada y verificar que cada reactivo contribuya efectivamente a la medición del constructo evaluado. De esta forma, se garantiza que los resultados reflejen diferencias reales en el dominio de los aprendizajes y no variaciones atribuibles a errores de construcción o funcionamiento inadecuado de los ítems.

Asimismo, la estimación de la confiabilidad como medida de consistencia interna proporciona evidencia sobre el grado en que la prueba mide de manera coherente los aprendizajes definidos en el marco de referencia. Habitualmente se le considera para optimizar la prueba como instrumento de medida, ya que brinda información para elegir los ítems que contribuyen a mejorar el test (Muñiz, Fidalgo, García-Cueto, Martínez & Moreno, 2005). En una evaluación diagnóstica, esta estabilidad es especialmente relevante, pues las decisiones pedagógicas derivadas de los resultados requieren sustento técnico suficiente para orientar diferentes intervenciones.

No obstante, si bien la **TCT** aporta herramientas valiosas para el análisis de la calidad técnica de la prueba, sus resultados se complementan con modelos de medición más robustos que permiten estimar el desempeño del estudiantado en una escala continua de habilidad. Esta combinación fortalece la precisión de la interpretación diagnóstica y asegura que la clasificación en niveles de desempeño responda tanto a criterios pedagógicos como a fundamentos psicométricos sólidos.

De este modo, la incorporación de la **Teoría Clásica de los Test** dentro del marco de referencia contribuye a la transparencia del proceso evaluativo y refuerza el compromiso institucional con una medición válida, confiable y orientada al fortalecimiento de los aprendizajes.

Teoría de respuesta a los ítems

La aplicación diagnóstica de las Pruebas Nacionales Estandarizadas por asignatura incorpora como fundamento psicométrico la **Teoría de Respuesta a los Ítems (TRI)**, específicamente el modelo de Rasch, con el propósito de estimar el desempeño del estudiantado en una escala continua de habilidad y fortalecer la precisión en la interpretación de los resultados.

La **TRI** se basa en dos postulados esenciales: 1) el rendimiento de una persona en un ítem se puede predecir o explicar por un conjunto de factores llamados rasgos o aptitudes latentes [*latent trait*]; y 2) la relación entre tal rendimiento y el conjunto de rasgos se puede describir por una función estrictamente creciente llamada función característica del ítem o curva característica del ítem, la cual especifica que a medida que el nivel del rasgo aumenta, la probabilidad de obtener la respuesta correcta para el ítem también aumenta (Hambleton & Cook, 1977; Hambleton, Swaminathan & Rogers, 1991). Estos rasgos suelen denominarse aptitud y se simbolizan con la letra θ (Martínez, 2005).

Las estimaciones de las puntuaciones de las personas en los rasgos latentes se utilizan para explicar la puntuación que tendrá la persona en un ítem o en una prueba completa, por lo que la variable independiente es el atributo o rasgo y la variable dependiente es la respuesta al ítem o a la prueba. De esta manera, a diferencia de la TCT, las puntuaciones empíricas son el resultado y no el fundamento del atributo (Martínez, 2005).

Dado el tipo de ítems de las Pruebas Nacionales Estandarizadas por asignatura, en particular, ítems de selección única, se trabajará con el modelo de **TRI** para puntuaciones dicotómicas conocido como **modelo de Rasch**.

El **modelo de Rasch** aporta herramientas de gran utilidad para la medición, ya que permite comprender con mayor claridad por qué las personas responden de determinada manera y cómo influyen las características de los ítems en ese comportamiento (Bond & Fox, 2001). En particular, ayuda a explicar la relación entre el nivel de habilidad del estudiantado y la dificultad de las tareas evaluadas, ofreciendo una base más sólida para interpretar los resultados. Su formulación fue planteada por Rasch en el año 1960, quien postuló que la respuesta a un ítem depende solo de la aptitud de la persona y de la dificultad del ítem.

A diferencia de enfoques basados únicamente en puntuaciones totales, el **modelo de Rasch** permite estimar la habilidad del estudiante y la dificultad de los ítems en una misma escala métrica. Esta característica posibilita comparar el nivel de desempeño de las personas con el grado de complejidad de las tareas evaluadas, proporcionando una base técnica sólida para interpretar qué tipo de aprendizajes han sido consolidados y cuáles requieren fortalecimiento.

Desde el punto de vista conceptual, el **modelo de Rasch** parte del principio de que la probabilidad de responder correctamente un ítem depende de la relación entre el nivel de habilidad del estudiante y la dificultad del ítem. Cuando la habilidad supera la dificultad, la probabilidad de acierto aumenta; cuando la dificultad es mayor que la habilidad estimada, la probabilidad disminuye. Esta relación estructurada permite ordenar tanto a estudiantes como a ítems en una misma escala, garantizando coherencia en la interpretación del desempeño.

En el contexto de una evaluación diagnóstica referida a criterios, esta propiedad resulta especialmente relevante. La interpretación de los resultados no se basa en comparaciones entre estudiantes, sino en la correspondencia entre el nivel de habilidad estimado y los aprendizajes definidos en el currículo oficial. La escala generada por el **modelo de Rasch** permite identificar con mayor precisión el punto en el continuo de aprendizaje en el que se ubica cada estudiante, favoreciendo una lectura pedagógica centrada en el dominio alcanzado y no en la posición relativa dentro de un grupo.

Asimismo, el **modelo de Rasch** ofrece ventajas técnicas fundamentales para la conformación de niveles de desempeño. Al situar los ítems en orden creciente de dificultad, es posible analizar qué tipo de tareas caracterizan distintos rangos de la escala. Esta organización facilita el establecimiento de puntos de corte mediante procedimientos estructurados de juicio experto, como el **método *Bookmark***, que vinculan la evidencia empírica con descriptores cualitativos del desempeño. De esta manera, los niveles de desempeño en las pruebas diagnósticas no se definen arbitrariamente, sino que se sustentan en la relación entre la dificultad de las tareas y el dominio esperado según el currículo oficial, cuya concreción se da en los programas de estudio vigentes para cada asignatura.

Otra fortaleza del **modelo de Rasch** radica en la estimación del error estándar de medición asociado a cada nivel de habilidad. A diferencia de enfoques que asumen un error uniforme para todos los estudiantes, este modelo permite estimar la precisión de la medición en distintos puntos de la escala. Esta característica resulta especialmente pertinente en la aplicación diagnóstica, pues contribuye a que la clasificación en niveles de desempeño se realice con mayor estabilidad y sustento técnico.

En coherencia con el enfoque referido a criterios adoptado por la prueba, el uso del **modelo de Rasch** refuerza la validez interpretativa de los resultados. La estimación de la habilidad no depende de la distribución del grupo evaluado, sino de la relación estructural entre el desempeño observado y la dificultad de las tareas alineadas al currículo. Esto asegura que los niveles de desempeño reflejen el grado de consolidación de los aprendizajes esperados y no posiciones relativas en una escala comparativa.

En consecuencia, la incorporación del **modelo de Rasch** en la interpretación de resultados de la aplicación diagnóstica fortalece la precisión de la medición, la coherencia entre currículo y evaluación, y la solidez técnica del establecimiento de niveles de desempeño. Su uso contribuye a que la interpretación de los resultados sea consistente con el propósito de la prueba y con el compromiso institucional de orientar decisiones pedagógicas basadas en evidencia rigurosa.

Niveles de **desempeño**

Los **niveles de desempeño** parten de un enfoque positivo de la evaluación, centrado en describir el grado de consolidación de los aprendizajes evidenciados y no simplemente en señalar carencias. En lugar de enfatizar lo que el estudiantado no sabe o no puede hacer, estos niveles se orientan a identificar aquello que tiene alta probabilidad de alcanzar con éxito, así como las oportunidades de fortalecimiento que permiten avanzar hacia mayores niveles de dominio. Desde esta perspectiva, la evaluación diagnóstica se concibe como un punto de partida para la mejora continua, promoviendo acciones pedagógicas que refuercen, amplíen y consoliden los aprendizajes en cada asignatura evaluada.

La definición de **niveles de desempeño** en la aplicación diagnóstica de las Pruebas Nacionales Estandarizadas por asignatura exige un fundamento psicométrico que garantice precisión en la estimación del desempeño y estabilidad en la clasificación del estudiantado. En este sentido, resulta fundamental considerar el tratamiento del error estándar de medición.

Desde la perspectiva de la **TCT**, el error de medición se concibe como una propiedad global de la prueba, asumiéndose que es constante para todas las personas evaluadas. Esta aproximación, aunque útil para estimar la confiabilidad general de los resultados, no permite identificar con precisión cómo varía la exactitud de la medición en distintos niveles de habilidad.

Por su parte, la **TRI**, y en particular el **modelo de Rasch**, permite estimar el error estándar asociado a cada nivel de habilidad. Esto significa que la precisión de la medición no es uniforme, sino que depende del punto de la escala en el que se ubique el estudiante. Esta característica es especialmente relevante cuando se establecen niveles de desempeño, pues la clasificación debe realizarse considerando la estabilidad de la estimación alrededor de los puntos de corte.

Al ubicar tanto a los ítems como a las personas en una misma escala continua, el modelo de Rasch permite identificar qué tareas caracterizan distintos **rangos del continuo de habilidad** y cuáles aportan mayor información en cada nivel. Esta organización de la escala constituye la base técnica para el establecimiento de niveles de desempeño mediante el **método *Bookmark***, procedimiento que combina evidencia empírica con juicio experto para determinar los puntos de corte que delimitan las categorías establecidas para los niveles de desempeño.

De esta manera, la clasificación del estudiantado no se basa únicamente en una puntuación total, sino en su ubicación en una escala de habilidad cuya estructura refleja la progresión de complejidad de los aprendizajes evaluados. El uso del error estándar asociado a cada estimación permite, además, valorar la estabilidad de dicha clasificación y reforzar la validez interpretativa de los niveles definidos.

Definición de los niveles de desempeño

Los **niveles de desempeño** corresponden a categorías que describen aquello que es capaz de hacer el estudiantado frente a lo que miden las pruebas. Estos niveles son esenciales para interpretar los resultados en términos del logro para la toma de decisiones por parte de los diferentes actores del sistema educativo en cada una de las asignaturas objeto de evaluación.

En otras palabras, los **niveles de desempeño** describen el grado de dominio evidenciado por la población estudiantil y, al mismo tiempo, orientan acciones pedagógicas diferenciadas para el

fortalecimiento del aprendizaje. Por ello, no deben interpretarse como etiquetas estáticas, sino como puntos de referencia dinámicos dentro de un proceso continuo de aprendizaje.

En coherencia con el propósito diagnóstico de las Pruebas Nacionales Estandarizadas por asignatura, los **niveles de desempeño** constituyen una referencia práctica para orientar acciones de mejora en los distintos ámbitos del proceso educativo. La información que brindan permite:

- Ajustar la mediación docente en el aula, identificando con mayor precisión las fortalezas del estudiantado y las áreas que requieren mayor profundización o refuerzo.
- Priorizar los aprendizajes y contenidos esenciales definidos en el currículo oficial, organizando el planeamiento con base en evidencias concretas del nivel de consolidación alcanzado.
- Orientar estrategias de acompañamiento y seguimiento académico por parte del personal docente y las familias, favoreciendo intervenciones oportunas y pertinentes.
- Fortalecer la gestión pedagógica a nivel institucional, facilitando la toma de decisiones por parte de los directores de centros educativos para implementar acciones de mejora sostenida.

De esta manera, los **niveles de desempeño** no se limitan a describir resultados, sino que se convierten en un insumo clave para la acción pedagógica y la mejora continua del aprendizaje.

Descriptores generales de los niveles de desempeño

Los **niveles de desempeño** se organizan de manera progresiva, lo que implica que un estudiante ubicado en un determinado nivel tiene alta probabilidad de responder correctamente los ítems de ese nivel, así como aquellos correspondientes a niveles inferiores o de menor complejidad. Esta estructura o lógica jerárquica refleja la naturaleza acumulativa del aprendizaje y la progresión en la dificultad de los ítems dentro del continuo de habilidad.

Para las Pruebas Nacionales Estandarizadas por asignatura en su propósito diagnóstico se definen **tres niveles de desempeño:** básico, intermedio y avanzado.

Los **descriptores generales** de los niveles de desempeño describen cualitativamente el nivel de dominio evidenciado en los aprendizajes evaluados en cada asignatura. Estos descriptores integran conocimientos, habilidades y procesos cognitivos establecidos en los programas de estudio vigentes, señalando aquello que el estudiantado tiene alta probabilidad de realizar de acuerdo con su ubicación en el continuo de habilidad.

Nivel básico. Corresponde a un **dominio elemental** de los conocimientos y las habilidades de cada asignatura que se evalúa en las pruebas diagnósticas. Describe a la población estudiantil que:

- Resuelve tareas de baja complejidad, sobre todo cuando estas se presentan de manera directa o estructurada.
- Presenta dificultades para aplicar los conocimientos de forma consistente o para integrarlos en situaciones que requieren mayor análisis o articulación de ideas.
- Requiere acompañamiento sistemático, práctica guiada y realimentación continua para fortalecer los aprendizajes y avanzar hacia desempeños más integrados y estables.

Aunque el **nivel básico** denota un desempeño estudiantil poco satisfactorio, con brechas iniciales y necesidades de intervención directa, también brinda insumos para elaborar estrategias que guíen hacia el mejoramiento que posibilite avanzar, gradualmente, al nivel intermedio.

Nivel intermedio. Corresponde a un **dominio parcial** de los conocimientos y las habilidades de cada asignatura que se evalúa en las pruebas diagnósticas. Describe a la población estudiantil que:

- Evidencia dominio consistente de los aprendizajes evaluados, resolviendo tareas de complejidad moderada y aplicando conocimientos en situaciones variadas dentro de lo establecido en cada programa de estudio.
- Integra conceptos y procedimientos de manera funcional, aunque pueden presentar dificultades cuando las tareas exigen mayor abstracción, transferencia a contextos nuevos o articulación simultánea de múltiples procesos.
- Requiere oportunidades de profundización y práctica en situaciones de mayor demanda cognitiva para avanzar hacia un desempeño más integrado y autónomo.

El nivel intermedio denota un desempeño estudiantil en proceso de alcanzar el dominio esperado y, además, brinda insumos para elaborar estrategias que orienten el mejoramiento y posibiliten continuar progresando gradualmente hacia el nivel avanzado.

Nivel avanzado. Corresponde a un **dominio satisfactorio** de los conocimientos y las habilidades de cada asignatura que se evalúa en las pruebas diagnósticas. Evidencia una alta probabilidad de que su desempeño esté en concordancia con las afirmaciones propias del nivel, lo que permite profundizar en los conocimientos y habilidades de forma significativa. En este nivel, la población estudiantil:

- Evidencia dominio sólido y estable de los aprendizajes evaluados, resolviendo tareas de mayor complejidad con consistencia y precisión.
- Integra y aplica conocimientos y procedimientos en contextos diversos, mostrando capacidad de análisis, transferencia y articulación de múltiples procesos cuando la situación lo requiere.
- Se encuentra en condiciones de abordar tareas desafiantes dentro de lo establecido en los programas de estudio, en procura de consolidar un desempeño autónomo y coherente con el nivel más alto esperado en la evaluación diagnóstica.

Antes de proceder al establecimiento de puntos de corte para los niveles básico, intermedio y avanzado, se define un umbral mínimo de evidencia de desempeño. Este umbral tiene como propósito asegurar que la clasificación en niveles se realice únicamente cuando la información disponible permita una interpretación técnicamente sustentada.

Cuando el número de aciertos no alcanza el mínimo establecido, la evidencia resulta insuficiente para ubicar de manera estable al estudiante en el continuo de habilidad. En tales casos se asigna la categoría “insuficiente”, la cual no constituye un nivel de desempeño dentro del continuo, sino una condición que indica necesidad prioritaria de acompañamiento pedagógico y verificación formativa adicional.

Esta decisión responde a criterios de validez interpretativa y resguarda la coherencia del proceso, evitando asignar niveles de desempeño cuando el patrón de respuestas no ofrece base suficiente para ello.

Naturaleza técnica de la categoría "Insuficiente"

La categoría **Insuficiente** no constituye un nivel de desempeño en sentido progresivo (como básico, intermedio o avanzado), sino una clasificación técnica que indica que la evidencia empírica disponible no permite ubicar con precisión al estudiante en uno de los tres niveles establecidos.

Esta categoría agrupa a los estudiantes que presentan un nivel muy bajo de desempeño, evidenciado por un porcentaje reducido de aciertos en la prueba (en torno al 15% o 20%, o incluso inferior), así como a quienes no responden correctamente ningún ítem. Desde el punto de vista psicométrico, estos casos se caracterizan por aportar una cantidad limitada de información para la estimación de la habilidad, lo que incrementa el error estándar asociado a su ubicación en la escala. En coherencia con los criterios técnicos definidos para el establecimiento de los puntos de corte, se consideran dentro de esta categoría aquellos desempeños cuya evidencia empírica no alcanza el umbral mínimo requerido para garantizar una clasificación estable y válida bajo el modelo de Rasch aplicado.

El porcentaje específico que delimita esta categoría se determinará a partir de análisis empíricos de precisión de la estimación en la escala Rasch, considerando el comportamiento del error estándar de medición y la estabilidad de los parámetros de habilidad en los rangos inferiores de desempeño. De esta manera, el umbral no responde a un criterio arbitrario de porcentaje de aciertos, sino a un análisis técnico de la calidad de la información aportada por las respuestas observadas.

Un estudiante que se ubica en la categoría "**insuficiente**" puede evidenciar dificultad para resolver tareas estructuradas básicas, mostrar respuestas mayoritariamente asociadas a distractores vinculados a errores conceptuales, o bien, poseer un dominio excesivamente limitado de los aprendizajes evaluados.

En el contexto de la aplicación diagnóstica, la categoría "insuficiente" exige un acompañamiento continuo y seguimiento periódico de los estudiantes así clasificados, dado que requieren un reforzamiento sistemático de los contenidos y las habilidades fundamentales contenidas en los programas de estudio.

Esta categoría no debe interpretarse como etiqueta permanente, sino como señal de intervención temprana, con planes de apoyo académico en el centro educativo, coordinación entre docentes, orientación y directores, así como estrategias de acompañamiento regional.

Establecimiento de puntos de corte mediante el método *Bookmark*

El establecimiento de los puntos de corte que delimitan los niveles de desempeño se realiza mediante el **método *Bookmark*,** procedimiento ampliamente utilizado en evaluaciones referidas a criterios para vincular la estructura empírica de la medición con juicios expertos fundamentados en el currículo oficial. Este método permite definir límites entre niveles de manera sistemática, transparente y técnicamente sustentada.

El proceso se inicia con la calibración de los ítems mediante el **modelo de Rasch**, lo que permite estimar su dificultad en una escala continua de habilidad. Una vez calibrados, los ítems se ordenan desde los de menor hasta los de mayor dificultad, conformando un cuadernillo ordenado de ítems. Este ordenamiento refleja la progresión empírica de complejidad de las tareas evaluadas y constituye la base objetiva sobre la cual se realiza el juicio experto.

El ordenamiento no depende de la frecuencia de aciertos en una muestra específica, sino de la estimación de dificultad derivada del modelo de medición, lo que garantiza estabilidad técnica y coherencia interna en la secuencia de ítems.

El establecimiento de los puntos de corte se realiza mediante un panel de jueces expertos en la asignatura correspondiente. El panel se conforma considerando criterios como:

- Experiencia docente en el nivel evaluado
- Conocimiento profundo del currículo oficial
- Experiencia en evaluación de aprendizajes
- Representatividad territorial y contextual
- Capacidad de análisis técnico-pedagógico

Precisamente esas características forman parte del perfil del puesto de asesor nacional y asesor regional en cada asignatura. Adicionalmente, antes de iniciar el proceso, los jueces reciben explicaciones y rondas de práctica específica sobre:

- El modelo de medición utilizado
- La interpretación del continuo de habilidad
- El significado de los niveles de desempeño
- El procedimiento formal del método *Bookmark*

Esta fase garantiza que las decisiones se tomen bajo un marco conceptual compartido y con comprensión técnica adecuada.

El **método *Bookmark*** se desarrolla en rondas estructuradas, las cuales se describen a continuación.

Primera ronda: juicio individual. Cada juez analiza el cuadernillo ordenado y determina el punto en el cual se ubicaría un estudiante mínimamente competente para el nivel inmediatamente superior. Es decir, identifica el lugar a partir del cual ese estudiante tendría la probabilidad objetivo previamente definida de responder correctamente los ítems correspondientes a ese nivel, pero no necesariamente los de mayor complejidad.

Como cada juez marca de manera individual el lugar donde considera que debe establecerse el límite entre niveles, se recopilan las ubicaciones propuestas y se analizan las diferencias entre jueces. Se presentan estadísticas descriptivas (mediana, rango, dispersión) para visualizar el grado de convergencia inicial.

Segunda ronda: deliberación y ajuste. Los jueces discuten las diferencias observadas, revisan ítems cercanos a las ubicaciones propuestas y fundamentan sus decisiones en relación con:

- Los descriptores de nivel.
- Las afirmaciones y evidencias curriculares.
- La complejidad cognitiva de las tareas.

Tras la discusión, cada juez puede mantener o ajustar su decisión inicial.

Rondas adicionales. Si la dispersión entre decisiones continúa siendo amplia, se desarrollan rondas adicionales hasta alcanzar un grado razonable de convergencia. El objetivo no es la unanimidad absoluta, sino la consistencia argumentada y técnicamente defendible.

El **método *Bookmark*** parte del supuesto de que el punto de corte debe representar el nivel de habilidad en el cual un estudiante tiene alta probabilidad de éxito en las tareas correspondientes a un nivel, pero no necesariamente en las de nivel superior. Esta probabilidad objetivo se define previamente y es consistente para todos los niveles establecidos.

El uso de una probabilidad objetivo permite vincular la dificultad de los ítems con el desempeño esperado, garantizando que los cortes no se basen únicamente en percepciones subjetivas, sino en la relación estructural entre habilidad y dificultad definida por el modelo de Rasch.

Para asegurar la calidad técnica del procedimiento, se implementan mecanismos de control tales como:

- Registro documentado de todas las decisiones.
- Análisis de dispersión entre jueces.
- Revisión de coherencia entre cortes y descriptores de nivel.
- Verificación de que los ítems cercanos a los puntos de corte representen adecuadamente las habilidades descritas.

Asimismo, se analiza la estabilidad de los cortes frente a la distribución empírica del desempeño, sin que esta determine la decisión.

Una vez definidos los puntos de corte, se realiza un análisis posterior que incluye:

- Revisión de la distribución del estudiantado en los niveles resultantes.
- Evaluación de la coherencia entre la clasificación obtenida y las expectativas curriculares.
- Análisis del error estándar de medición en torno a los puntos de corte para valorar la estabilidad de la clasificación.

Este proceso no modifica automáticamente los cortes definidos, pero permite verificar que las decisiones sean técnica y pedagógicamente consistentes.

El **método *Bookmark*** resulta particularmente pertinente en una evaluación diagnóstica, ya que permite definir niveles de desempeño en función del dominio curricular esperado en cada asignatura y no de la posición relativa del estudiantado dentro del grupo evaluado. La clasificación se fundamenta en logros de aprendizaje y no en porcentajes de aprobación ni en comparaciones normativas.

De este modo, los niveles establecidos describen grados progresivos de consolidación de aprendizajes y ofrecen información útil para orientar la mejora pedagógica en cada asignatura, manteniendo coherencia entre información psicométrica, fundamento curricular e interpretación formativa de los resultados.

Ahora bien, dado que toda medición incorpora un margen de error, la clasificación del estudiantado

en niveles de desempeño considera la proximidad de las estimaciones de habilidad a los puntos de corte establecidos. El modelo de Rasch permite estimar el error estándar asociado a cada medición individual, lo cual ofrece información sobre la precisión de la ubicación en el continuo de habilidad.

En los casos en que la estimación de un estudiante se ubica muy próxima a un punto de corte, se reconoce que pequeñas variaciones dentro del margen de error podrían situarlo en uno u otro nivel adyacente. Por esta razón, la interpretación de la clasificación debe realizarse con criterio pedagógico y no como una determinación rígida. El propósito del nivel asignado es orientar la acción educativa y no establecer juicios definitivos. Esta consideración refuerza la naturaleza diagnóstica del proceso y evita interpretaciones absolutas basadas en diferencias marginales.

Asimismo, el análisis posterior al establecimiento de los puntos de corte incluye la revisión de la estabilidad de la clasificación alrededor de dichos límites, verificando que los cortes no generen concentraciones artificiales o desplazamientos abruptos que carezcan de fundamento curricular.

El establecimiento de puntos de corte mediante el método *Bookmark* se desarrolla bajo criterios de transparencia y documentación rigurosa. Todas las etapas del proceso (conformación del panel, capacitación, rondas de juicio, deliberaciones, ajustes y validación posterior) quedan registradas formalmente, garantizando trazabilidad en la toma de decisiones.

Asimismo, se documentan los fundamentos curriculares utilizados por los jueces, las razones que sustentan los ajustes realizados entre rondas y los análisis de consistencia efectuados posteriormente. Esta sistematicidad permite que el proceso sea replicable, auditable y defendible técnicamente ante cualquier instancia académica o institucional.

La transparencia del procedimiento refuerza la legitimidad de los niveles de desempeño definidos y evidencia que estos no derivan de decisiones arbitrarias, sino de un proceso estructurado que integra medición psicométrica, juicio experto disciplinar y coherencia curricular.

Descriptores específicos por asignatura

Además de los descriptores generales que caracterizan los niveles básico, intermedio y avanzado, se elaboran **descriptores específicos** para cada asignatura evaluada. Estos descriptores detallan, con mayor precisión disciplinar, los conocimientos, habilidades y procesos cognitivos que el estudiantado tiene alta probabilidad de demostrar según su ubicación en el continuo de habilidad.

Los **descriptores específicos** se formulan en estricta congruencia con el marco de especificaciones de cada prueba. En este sentido, se fundamentan en las afirmaciones y evidencias definidas previamente, así como en la distribución de contenidos y procesos cognitivos establecidos para la evaluación. Esta coherencia asegura que los niveles de desempeño no se interpreten de manera genérica, sino en relación directa con los aprendizajes efectivamente evaluados en cada asignatura.

El proceso de elaboración de los **descriptores específicos** integra tres elementos centrales: la estructura curricular oficial, el análisis del dominio desarrollado en el marco de referencia y la organización empírica de los ítems según su dificultad. De esta manera, los descriptores no constituyen enunciados abstractos, sino síntesis interpretativas construidas a partir de la evidencia que respalda el establecimiento de los puntos de corte mediante el método *Bookmark*.

Asimismo, los **descriptores específicos** cumplen una función clave en la comunicación de resultados. Constituyen el insumo técnico para la elaboración de los informes diagnósticos, tanto a nivel individual, por centro educativo, a nivel regional y nacional. En dichos informes, los descriptores permiten traducir la estimación cuantitativa del desempeño en información pedagógicamente significativa, facilitando la comprensión de aquello que el estudiantado puede realizar con consistencia y de los aspectos que requieren fortalecimiento.

En consecuencia, la articulación entre marco de especificaciones, modelo de medición, establecimiento de niveles y descriptores específicos garantiza coherencia interna en el sistema evaluativo a nivel macro. Esta integración asegura que la interpretación diagnóstica de los resultados esté alineada con el currículo oficial, con la estructura técnica de la prueba y con el propósito que orienta la evaluación.

Indicador de Desempeño Estandarizado de Aprendizajes (IDEA-250)

Con el propósito de fortalecer la interpretación diagnóstica de los resultados y asegurar coherencia entre la medición y su comunicación educativa, se presenta el **Indicador de Desempeño Estandarizado de Aprendizajes (IDEA-250)** como la métrica de reporte en cada asignatura evaluada. Este indicador constituye el eje cuantitativo que respalda la clasificación en niveles de desempeño y orienta la lectura técnica y pedagógica del grado de consolidación de los aprendizajes incluidos en la prueba diagnóstica de cada asignatura.

Propósito

A partir de la aplicación diagnóstica 2026, el desempeño del estudiantado en cada asignatura se reportará mediante el **Indicador de Desempeño Estandarizado de Aprendizajes (IDEA-250)**. Este indicador constituye la métrica de comunicación de resultados en el marco de la evaluación diagnóstica y se fundamenta en estimaciones obtenidas a través del modelo de Rasch.

El **IDEA-250** es una estimación estandarizada del desempeño que expresa la ubicación del estudiante en un continuo de habilidad, considerando la dificultad de las tareas evaluadas y la consistencia de su patrón de respuestas. Su finalidad es describir el grado de consolidación de los aprendizajes establecidos en el currículo oficial de cada asignatura. No es una nota ni un porcentaje de aciertos.

Las estimaciones originales del modelo de Rasch se expresan en *logits*, unidad técnica propia del modelo. Con el propósito de facilitar la comprensión y comunicación de resultados, dichas estimaciones se transforman linealmente a una escala con media fijada en 250 y desviación estándar fijada en 40 puntos. Esta transformación preserva el orden relativo de los desempeños, las distancias entre estimaciones y el error estándar asociado a cada medición.

Fundamentación psicométrica

La fijación de la media y la desviación estándar del IDEA-250 responde a una convención de reporte. Sus puntajes son interpretables dentro de cada asignatura y de la métrica construida para ese dominio.

La **escala IDEA-250** mantiene intacta la estructura técnica de la medición; únicamente modifica la unidad de reporte para hacerla más interpretable en el ámbito educativo. En consecuencia, cualquier análisis realizado sobre la escala transformada conserva la validez de la medición original.

Justificación educativa

La adopción del **IDEA-250** responde a una decisión orientada a fortalecer el carácter diagnóstico de la evaluación. La escala tradicional de 0 a 100 puntos se encuentra culturalmente asociada a procesos de aprobación o reprobación y a la noción de nota mínima para "pasar", lo cual no es coherente con el propósito de esta aplicación.

El **IDEA-250** desvincula la interpretación del desempeño de la lógica de calificación sumativa y centra la atención en los niveles de dominio alcanzados. De esta manera, la pregunta deja de ser "¿cuánto obtuvo?" para convertirse en "¿qué nivel de consolidación de aprendizajes evidencia?". Este cambio favorece una lectura educativa orientada a la mejora continua y no a la certificación.

Relación con los niveles de desempeño

El **IDEA-250** constituye el soporte cuantitativo para la clasificación en los niveles básico, intermedio y avanzado, establecidos mediante el **método *Bookmark***. El puntaje permite ubicar al estudiante en el continuo de habilidad, mientras que el nivel ofrece la interpretación cualitativa correspondiente en términos de dominio curricular.

El indicador y los niveles funcionan de manera complementaria, esto es, el puntaje proporciona precisión técnica y el nivel aporta significado educativo. Esta articulación asegura coherencia entre medición e interpretación curricular.

Aplicación por asignatura

El **IDEA-250** se aplica de manera uniforme en todas las asignaturas evaluadas, garantizando consistencia en la estructura de reporte. No obstante, cada resultado debe interpretarse exclusivamente dentro del marco curricular específico de la asignatura correspondiente. La uniformidad de la escala no implica comparabilidad directa entre áreas disciplinares, dado que cada prueba evalúa aprendizajes propios de su campo de conocimiento, en relación directa con los programas de estudio vigentes en cada asignatura.

Finalmente, es importante señalar que el **IDEA-250** no debe interpretarse como:

- Una nota de aprobación o reprobación.
- Un porcentaje directo de respuestas correctas.
- Una clasificación comparativa entre estudiantes.

Se trata de un indicador referido a criterios, cuya interpretación se realiza en función del dominio de aprendizajes definidos curricularmente y no en relación con la posición relativa dentro del grupo evaluado.

Evidencias de **validez y confiabilidad**

De acuerdo con Ravela (2006), el concepto de validez se refiere al grado en que los juicios de valor que se formulan de los resultados de una evaluación estén debidamente sustentados y estén efectivamente relacionados con el "referente" definido para la evaluación; mientras que la confiabilidad se refiere a la precisión de las medidas y de la evidencia empírica empleada en la evaluación.

Como parte de las evidencias de validez de las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico, posteriormente al análisis del dominio, se elabora una tabla de especificaciones, la cual operacionaliza el constructo y orienta con claridad el proceso evaluativo. Estas tablas fueron revisadas por asesores regionales y nacionales del Ministerio de Educación Pública y pueden ser consultadas en los marcos de especificaciones de las Pruebas Nacionales Estandarizadas por asignatura con propósito diagnóstico.

Estudio piloto

Como parte del aseguramiento de la calidad técnica de las Pruebas Nacionales Estandarizadas por asignatura, los conjuntos de ítems son sometidos a un estudio piloto previo a su inclusión en la aplicación diagnóstica y del resto de pruebas que se generan a lo largo del año. Este pilotaje constituye una condición esencial para la generación de evidencias de validez, en tanto permite examinar empíricamente el funcionamiento de los ítems y su coherencia con el constructo definido en el marco de referencia y en el marco de especificaciones de cada asignatura.

Desde la perspectiva de la validez como argumento sustentado en evidencia, el estudio piloto aporta información relevante en distintas dimensiones. En primer lugar, ofrece evidencias relacionadas con la estructura interna de la prueba, al analizar el comportamiento psicométrico de los ítems mediante indicadores como dificultad, discriminación, consistencia interna y ajuste al modelo de medición adoptado. Estos análisis permiten verificar que los ítems contribuyan adecuadamente a la estimación del continuo de habilidad y que el instrumento funcione de manera coherente con el constructo evaluado.

En segundo lugar, el pilotaje aporta evidencias sobre los procesos de respuesta, al examinar el funcionamiento de los distractores, la claridad de los enunciados y la plausibilidad de las opciones de respuesta. El análisis de patrones de selección y de funcionamiento diferencial permite identificar posibles ambigüedades, sesgos o interpretaciones no previstas, fortaleciendo así la validez interpretativa de los resultados.

Asimismo, el **estudio piloto** permite contrastar la congruencia entre los resultados empíricos y las afirmaciones y evidencias establecidas en el marco de especificaciones. Esta verificación asegura que la dificultad observada de los ítems sea coherente con la progresión de complejidad prevista y que las tareas evaluadas representen adecuadamente los aprendizajes definidos en los programas de estudio de cada asignatura.

En conjunto, el proceso de pilotaje no solo depura técnicamente los ítems, sino que fortalece el argumento de validez que sustenta la interpretación de los resultados diagnósticos. Al integrar evidencia empírica, fundamentación curricular y revisión experta, se garantiza que las inferencias derivadas de la prueba sean consistentes, pertinentes y técnicamente defendibles dentro del sistema nacional de evaluación.

Finalmente, cabe destacar que los resultados del **estudio piloto** tienen carácter estrictamente técnico y se utilizan exclusivamente para fines internos de análisis, ajuste y aseguramiento de la calidad de la prueba. Su propósito no es generar reportes de desempeño ni establecer conclusiones sobre el estudiantado participante, sino evaluar el funcionamiento psicométrico de los ítems y fortalecer las evidencias de validez que respaldan la aplicación oficial. En resguardo de la integridad del proceso evaluativo y para evitar interpretaciones indebidas de información preliminar, los resultados del pilotaje no se divulgan públicamente. Esta práctica responde a criterios de rigor técnico, ética evaluativa y protección del uso adecuado de los instrumentos dentro del sistema nacional de evaluación a nivel macro.

Juzgamiento de ítems

En el marco de la construcción técnica de las pruebas, los ítems son sometidos a un proceso formal de juzgamiento por parte de asesores nacionales y regionales de cada asignatura. Este procedimiento constituye una fuente fundamental de evidencias de validez basadas en el contenido, al garantizar que las tareas evaluadas representen de manera adecuada los aprendizajes establecidos en el currículo oficial.

Durante el juzgamiento se analiza la correspondencia de cada ítem con las afirmaciones y evidencias definidas en el marco de especificaciones, la pertinencia del nivel de complejidad cognitiva, la claridad y precisión del enunciado, así como la coherencia y plausibilidad de las opciones de respuesta. Asimismo, se valoran posibles sesgos lingüísticos, culturales o contextuales que puedan afectar la interpretación de la tarea. Este proceso fortalece la representatividad y pertinencia del contenido evaluado, asegurando que las pruebas respondan de manera consistente al constructo definido y que las inferencias derivadas de sus resultados estén sustentadas en una revisión técnica rigurosa.

Además, se cuenta con procesos internos y externos de juzgamiento de ítems por parte de expertos de cada una de las asignaturas. Las observaciones emanadas en los **procesos de juzgamiento** son posteriormente valoradas por los equipos técnicos para la incorporación en las versiones finales de los ítems.

En la figura 2, se muestra la tabla utilizada para el juzgamiento de ítems que se le brinda a los equipos de asesores técnicos (nacionales y regionales) con el fin de registrar las evidencias y observaciones al respecto.

Figura 2

Tabla para el juzgamiento de ítems de las Pruebas Nacionales Estandarizadas por asignatura 2026

MINISTERIO DE EDUCACIÓN PÚBLICA | GOBIERNO DE COSTA RICA | DGEC Dirección de Gestión y Evaluación de la Calidad

PRUEBAS NACIONALES ESTANDARIZADAS
JUZGAMIENTO DE LA CALIDAD TÉCNICA DEL ÍTEM

Nombre del juez: _____ **Fecha:** ____/ ____/ 2026

Habilidades de la Política Curricular			
Pensamiento sistémico ○	Pensamiento crítico ○	Resolución de problemas ○	
Bloque de conocimientos medido	**Afirmación medida**	**Evidencias**	**Respuesta correcta (clave)**
Nº _____	**Nº** _____	**Nº** _____	**A** ○ **B** ○ **C** ○ **D** ○

Contenido	SÍ	NO
Uso correcto de la gramática y la ortografía.	○	○
La redacción del enunciado es coherente y afirmativa.	○	○
El vocabulario es adecuado al desarrollo cognitivo del estudiantado.	○	○
La información es relevante para el planteamiento y la solución del ítem.	○	○

Opciones de respuesta	SÍ	NO
La clave es incuestionable.	○	○
Las opciones de respuesta son excluyentes entre sí.	○	○
La redacción de las opciones de respuesta es afirmativa.	○	○
Existe concordancia entre el enunciado y las opciones de respuesta.	○	○

Sensibilidad

El ítem presenta un

lenguaje no inclusivo ○ vocabulario técnico no pertinente al componente ○

El ítem presenta implicaciones SÍ ○ NO ○

políticas ○ de género ○ ideológicas ○ xenofóbicas ○ propagandistas ○

religiosas ○ deportivas ○ etnofóbicas ○ homofóbicas ○

Imágenes* SÍ ○ NO ○		
	SÍ	NO
La resolución de la imagen es óptima.	○	○
La imagen es pertinente para la solución del ítem.	○	○
La calidad de la imagen es clara (color, luz y texturas).	○	○
La imagen complementa la información del enunciado.	○	○

***Imágenes:** símbolos, gráficos, dibujos, esquemas, figuras, otros.

Confiabilidad de los resultados en la aplicación diagnóstica

La **confiabilidad** constituye un requisito esencial para garantizar que las interpretaciones derivadas de la aplicación diagnóstica sean estables y técnicamente sustentadas. En este contexto, la **confiabilidad** se entiende como el grado en que los resultados reflejan consistentemente el desempeño del estudiantado en relación con los aprendizajes evaluados, minimizando el efecto del error de medición.

Desde la perspectiva de la TCT, la consistencia interna de la prueba se estima mediante el coeficiente alfa de Cronbach. Este indicador permite valorar el grado en que los ítems que conforman la prueba funcionan de manera coherente entre sí al medir un mismo constructo. En una evaluación diagnóstica por asignatura, un valor adecuado de alfa sugiere que las tareas incluidas comparten una base común de contenido y que el puntaje obtenido constituye una medida estable del desempeño dentro del dominio evaluado. No obstante, el alfa debe interpretarse considerando la naturaleza del instrumento, la cantidad de ítems y la amplitud del contenido evaluado, evitando su uso como único criterio de calidad técnica.

Complementariamente, desde el modelo de Rasch, la confiabilidad se aborda mediante indicadores como la confiabilidad de personas y la confiabilidad de ítems, así como el análisis del error estándar de medición asociado a cada estimación individual. A diferencia de la TCT, el modelo de Rasch permite estimar el error de manera específica a lo largo del continuo de habilidad, lo que aporta mayor precisión para interpretar la estabilidad de la clasificación en niveles de desempeño. Esto resulta especialmente relevante en una aplicación diagnóstica, donde la ubicación del estudiantado en los niveles básico, intermedio o avanzado debe sustentarse en estimaciones técnicamente consistentes.

Desde esta perspectiva, garantizar la **confiabilidad** implica no solo alcanzar niveles adecuados de consistencia interna, sino también verificar el ajuste de los ítems al modelo de medición, asegurar la unidimensionalidad del constructo evaluado y analizar la estabilidad de las estimaciones en torno a los puntos de corte establecidos. Asimismo, es fundamental examinar el comportamiento de los ítems para evitar que aquellos con bajo poder discriminativo o con funcionamiento inadecuado introduzcan error innecesario en la medición.

Tipos de ítems en las **Pruebas Nacionales Estandarizadas**

Las Pruebas Nacionales Estandarizadas por asignatura solamente incluyen ítems de selección única, en los que la respuesta correcta se denomina "clave" y las otras opciones se conocen como "distractores".

En particular, los ítems de selección única son muy útiles en la medición de conocimientos que requieran distintas demandas cognitivas, aunque claramente no son los únicos existentes. A pesar de que su elaboración no es sencilla, permiten dicotomizar la respuesta en términos de acierto/no acierto para efectos de puntuación en una prueba, lo que facilita la automatización en los procesos de calificación, sobre todo en aplicaciones de gran escala.

Es común que se les asocie con ítems que se respondan tan solo evocando algún término o concepto, sin embargo, realmente permiten medir la aplicación de conocimientos, habilidades y otras destrezas en la resolución de tareas complejas (Haladyna & Rodríguez, 2013).

Cantidad de **opciones**

A partir del año 2026, la aplicación diagnóstica de las Pruebas Nacionales Estandarizadas por asignatura estará conformada por **ítems de selección única con cuatro opciones de respuesta**: una alternativa correcta (clave) y tres distractores plausibles.

Esta modificación forma parte de un proceso de fortalecimiento técnico orientado a mejorar la precisión de la medición y la calidad de la información que se brinda a docentes, centros educativos y autoridades para la toma de decisiones pedagógicas al inicio del ciclo lectivo.

El aumento de tres a cuatro opciones reduce la probabilidad de acierto por azar de aproximadamente 33% a 25%. Esta disminución es relevante en una aplicación diagnóstica, pues permite que el desempeño observado refleje con mayor fidelidad el nivel real de dominio del estudiantado.

Cuando la probabilidad de éxito aleatorio es menor:

- Se reduce la influencia de factores no relacionados con el aprendizaje.
- Se obtiene una estimación más estable del nivel de habilidad.
- Se mejora la identificación de estudiantes con dominio incipiente o parcial.

En el contexto diagnóstico, esta mayor precisión es fundamental para detectar brechas reales de aprendizaje y orientar oportunamente la mediación docente.

Por otra parte, la incorporación de un cuarto distractor no implica un aumento artificial de la dificultad conceptual de los ítems.

Desde la Teoría Clásica de los Test, el índice de dificultad refleja la proporción de estudiantes que responde correctamente. Al reducirse el acierto por azar, el valor observado de dificultad se depura del componente aleatorio y se aproxima más al dominio efectivo del contenido evaluado.

En consecuencia:

- El ítem no se vuelve más complejo en términos curriculares.
- Se reduce la probabilidad de respuestas correctas sin dominio real.
- La dificultad estimada representa de manera más precisa el aprendizaje alcanzado.

Esto es especialmente importante en una evaluación diagnóstica, donde el objetivo es conocer con claridad el punto de partida del estudiantado.

Los distractores deben ser plausibles, esto es, que representen un error auténtico y frecuente en estudiantes con dominio parcial. La inclusión de un tercer distractor permite capturar una mayor diversidad de errores típicos, lo que fortalece la función diagnóstica del ítem en cada asignatura. Cuando se cuenta con tres distractores plausibles el ítem discrimina de mejor manera entre grupos alto y bajo en relación con la mediana.

Por otra parte, en el modelo de Rasch, cada ítem aporta información en una zona específica del continuo de habilidad. De esta manera, al reducir el azar y mejorar la calidad de los distractores se incrementa la calidad de la información del ítem, se mejora la precisión en la estimación de habilidad y, en consecuencia, se fortalece la estabilidad de la clasificación en niveles de desempeño.

Esto es especialmente relevante en evaluación diagnóstica referida a criterios, donde la precisión en torno a los puntos de corte es fundamental.

Formatos de ítems de **selección única**

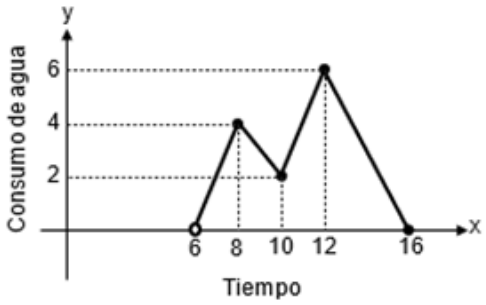
La elaboración de ítems requiere de personas expertas en el área de conocimientos por medir, así como experiencia en el ámbito educativo. Este es un proceso que también involucra creatividad y claridad en cuáles son los conocimientos y las habilidades que se emplean en la resolución de tareas de distinta índole y complejidad en cada una de las asignaturas que son objeto de evaluación.

Los ítems de selección única pueden presentarse de forma individual o como multirreactivos; este último formato permite evaluar diversos conocimientos y habilidades interrelacionados que se desprenden de la información que se muestra al inicio del ítem, aunque cada pregunta se resuelve y evalúa de manera independiente.

En el ejemplo 2, se presenta un ejemplo de multirreactivo en la asignatura de "Matemáticas" para secundaria.

Para responder los ítems 12 y 13 considere la siguiente información:

La siguiente representación gráfica corresponde a la cantidad de agua potable, en metros cúbicos, que se consumió en una institución en función del tiempo "x", en horas de un día, con $6 < x \leq 16$:



12) La cantidad de agua que se consumió, en esa institución, disminuyó entre las

A) 6 h y 8 h.
B) 8 h y 10 h.
C) 10 h y 12 h.
D) 11 h y 12 h.

13) ¿A qué hora de ese día se consumió la mayor cantidad de agua en esa institución?

A) A las 6 h
B) A las 8 h
C) A las 12 h
D) A las 16 h

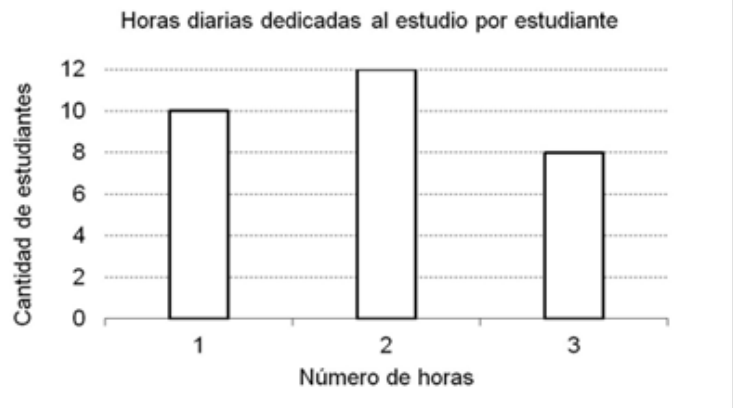
Aunque existen diversos formatos para los ítems de selección única, en las Pruebas Nacionales Estandarizadas se utilizarán los dos siguientes:

- **Formato interrogativo:** Se presenta una situación y una pregunta con base en ella. Permiten medir tanto la comprensión como la aplicación de conocimientos, habilidades y otras destrezas para seleccionar la respuesta correcta.

- **Formato de encabezado incompleto:** Se presenta, en el encabezado, una oración incompleta; las opciones completan el encabezado gramaticalmente (debe existir concordancia entre género y número).

Asimismo, en las Pruebas Nacionales Estandarizadas se emplearán algunas formas complejas de ítems de selección única expuestos en Aiken (2003) y en Haladyna y Rodríguez (2013). En el ejemplo 3, mostramos un ítem de selección única en que las opciones presentan una justificación para la respuesta.

Una maestra recolectó datos, por medio de una encuesta, sobre el número de horas diarias que dedican al estudio los 30 estudiantes de una sección. A continuación, se muestra el gráfico que elaboró la maestra a partir de los datos recolectados:



La maestra analizó la información anterior y afirmó que 3 corresponde al mayor número de horas diarias que dedican al estudio esos estudiantes. ¿Está usted de acuerdo con esa afirmación?

A) Sí, porque corresponde al valor máximo.
B) No, porque el valor máximo es 12.
C) Sí, porque corresponde a la moda.
D) No, porque 3 representa la menor cantidad de estudiantes.

Diseño Universal de **Evaluación**

Este es un enfoque que, en pruebas de gran escala, permite una flexibilización del diseño y desarrollo que favorezca la participación del mayor número posible de personas estudiantes, así como para hacer inferencias válidas sobre su desempeño en este tipo de evaluaciones (Thompson et al., 2002).

Thompson y Thurlow (2002), así como Thompson et al. (2002) y Hanna (2005) han propuesto una serie de principios para el diseño universal de evaluación, los cuales se describen a continuación:

- **Evaluación inclusiva de la población:** Desde el diseño de la prueba, se debe tener como norte la inclusión de toda la población estudiantil, tomando en consideración el amplio rango de condiciones presentes en dicha población.

- **Definición precisa de lo que se va a medir:** Se debe establecer con precisión lo que se va a medir en la prueba, de modo que se eliminen todas las barreras cognitivas, sensoriales, emocionales y físicas irrelevantes.

- **Accesibilidad y eliminación del sesgo en los ítems:** Los principios que se refieren a la accesibilidad deben incorporarse al diseño de ítems desde el inicio del proceso y elaborar revisiones que garanticen que se mantenga la calidad en los ítems, eliminando toda forma de sesgo (ventajas o desventajas en la presentación o el contenido del ítem que puedan invalidar su aporte a la puntuación de una prueba).

- **Adaptaciones:** El diseño de los ítems debe permitir las adaptaciones que se requieran a fin de reducir las amenazas contra las evidencias de validez y comparabilidad de puntuaciones. Por ejemplo, prueba transcrita a sistema Braille.

- **Instrucciones y procedimientos sencillos, claros e intuitivos:** Todas las instrucciones y los procedimientos son sencillos, claros y se presentan en un lenguaje comprensible.

- **Máxima legibilidad y comprensibilidad:** Se debe procurar, en la elaboración de ítems, el uso de un lenguaje sencillo (por ejemplo, organización del texto, reducción de la longitud de frases y cantidad de palabras consideradas como difíciles) para que el texto sea legible y comprensible. Este mismo principio aplica para figuras y tablas, así como formatos de respuesta presentes en un ítem.

Para las Pruebas Nacionales Estandarizadas por asignatura en su aplicación diagnóstica se han seguido los principios del diseño universal de evaluación, como una forma de garantizar mayor accesibilidad de la población estudiantil a esta prueba. Asimismo, se elaboró y publicó el documento denominado **"Orientaciones para el abordaje de los apoyos educativos y los ajustes razonables en las pruebas nacionales"** mediante el cual se guía a las personas usuarias para que realicen la gestión de los apoyos educativos y ajustes razonables de manera tal que se minimicen las barreras para el aprendizaje y la participación de las personas estudiantes.

Referencias **bibliográficas**

Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Allyn and Bacon.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.

Haladyna, T.M., & Rodríguez, M.C. (2013). *Developing and Validating Test Items*. Routledge.

Haladyna, T.M., Downing, S.M., & Rodríguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Hambleton, R. K., & Cook, L. L. (1977). *Latent trait models and their use in the analysis of educational test data*. Journal of Educational Measurement, *14*(2), 75-96. https://doi.org/10.1111/j.1745-3984.1977.tb00030.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hanna, E. (2005). *Inclusive Design for Maximum Accessibility: A Practical Approach to Universal Design*. Pearson Educational Measurement.

Martínez Arias, M. R. (2005). *Psicometría: Teoría de los tests psicológicos y educativos*

Ministerio de Educación Pública de Costa Rica. (2017). *Política Educativa "La persona: Centro del proceso educativo y sujeto transformador de la sociedad"*.

Ministerio de Educación Pública de Costa Rica. (2015). *Política Curricular "Educar para una nueva ciudadanía".*

Ministerio de Educación Pública de Costa Rica. (2025). *Orientaciones para el abordaje de los apoyos educativos y los ajustes razonables en las pruebas nacionales.*

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered design: Layers, Structures, and Terminology*. SRI International.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence Centered Design. Educational Testing Service.*

Mislevy, R. J., Haertel, G., Riconscente, M., Wise Rutstein, D., & Ziker, C. (2017). *Assessing Model-Based Reasoning using Evidence- Centered Design: A Suite of Research-Based Design Patterns.* Springer.

Muñiz, J., Fidalgo, A. M., García-Cueto, E., Martínez, R., & Moreno, R. (2005). *Análisis de los ítems*. La Muralla.

Muñiz, J. (2018). *Introducción a la Psicometría: Teoría Clásica y TRI.* Ediciones Pirámide.

Parshall, C.G., Harmes, J.C., Davey, T., & Pashley, P.J. (2009). Innovative Items for Computerized Testing En W. van der Linden, & C. Glas (eds). *Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences* (pp. 215-230). Springer.

Ravela, P. (2006). *Fichas didácticas para comprender las evaluaciones educativas.* PREAL.

Rodríguez, M.C. (2005). *Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research*. *Educational Measurement: Issues and Practice, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rodríguez Frías, M. B., y Flotts de los Hoyos, M. P. (2019). *Definición del referente de la evaluación y desarrollo del marco de especificaciones. Cuadernillo técnico de evaluación educativa 3*. Centro de Medición MIDE UC e Instituto Nacional para la Evaluación de la Educación INEE.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Policy14.html

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79-87, http://dx.doi.org/10.1016/j.pse.2014.11.003

Autoridades **ministeriales**

José Leonardo Sánchez Hernández
Ministro de Educación Pública

Guiselle Alpízar Elizondo
Viceministra Académica

Alejandra Gutiérrez Vargas
Viceministra de Planificación Institucional y Coordinación Regional

Sofía Ramírez González
Viceministra Administrativa

Álvaro Artavia Medrano
Director, Dirección de Gestión y Evaluación de la Calidad

Ana Carvajal Granados
Subdirectora, Dirección de Gestión y Evaluación de la Calidad

San José, Costa Rica
Marzo 2026

Equipo técnico

Ciencias
Ramón Montoya Jiménez
José Fabio Gámez Romero
Johanna Segura Solano

Estudios Sociales
Ana Balbina Flores Cornejo
Humberto Hernández Rodríguez

Español
Wilfredo Acevedo Mojica
Jorge Fonseca Rojas
Vanessa Zárate Montero

Educación Cívica
Noelia Alvarado González

Matemáticas
Rafael González Palacios
Omar Guzmán Alvarado
Gerardo Murillo Vega

Apoyos educativos
Luis Carlos Rodríguez León