INSTITUTO NACIONAL PARA LA EVALUACION DE LA EDUCACION

CRITERIOS técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la definición de las listas de prelación de los Concursos de oposición para el ingreso al Servicio Profesional Docente en Educación Básica y Educación Media Superior para el ciclo escolar 2019-2020.

Al margen un logotipo, que dice: Instituto Nacional para la Evaluación de la Educación.- México.

CRITERIOS TÉCNICOS Y DE PROCEDIMIENTO PARA EL ANÁLISIS DE LOS INSTRUMENTOS DE EVALUACIÓN, EL PROCESO DE CALIFICACIÓN Y LA DEFINICIÓN DE LAS LISTAS DE PRELACIÓN DE LOS CONCURSOS DE OPOSICIÓN PARA EL INGRESO AL SERVICIO PROFESIONAL DOCENTE EN EDUCACIÓN BÁSICA Y EDUCACIÓN MEDIA SUPERIOR PARA EL CICLO ESCOLAR 2019-2020

Con fundamento en los Lineamientos para llevar a cabo la evaluación para el ingreso al Servicio Profesional Docente en Educación Básica y Media Superior para el ciclo escolar 2019-2020 (LINEE-01-2019), la Junta de Gobierno aprueba emitir los siguientes Criterios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la definición de las listas de prelación de los Concursos de oposición para el ingreso al Servicio Profesional Docente en Educación Básica y Educación Media Superior para el ciclo escolar 2019-2020.

Los presentes Criterios técnicos y de procedimiento consideran el uso de los datos recabados una vez que se ha llevado a cabo la aplicación de los instrumentos que forman parte de la evaluación y tienen como finalidad establecer los referentes necesarios para garantizar la validez, confiabilidad y equidad de los resultados de los procesos de evaluación implicados en estos concursos. Su contenido se organiza en cinco apartados: 1) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 2) Procedimiento para el establecimiento de puntos de corte y estándares de desempeño; 3) Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos del proceso de evaluación; 4) Resultado del proceso de evaluación: conformación de los grupos de desempeño y 5) Integración de las listas de prelación.

En la parte final, se integra un Anexo técnico con información detallada de los aspectos que se consideran en los distintos apartados del documento.

Definición de términos

Para los efectos del presente documento, se emplean las siguientes definiciones:

I. **Alto impacto:** Se entiende que una evaluación es de alto impacto cuando sus resultados tienen consecuencias importantes para las personas o las instituciones; por ejemplo, los procesos de admisión o certificación.

II. **Calificación:** Proceso de asignación de una puntuación o nivel de desempeño logrado a partir de los resultados de una medición.

III. **Confiabilidad:** Cualidad de las mediciones obtenidas con un instrumento, que se caracterizan por ser consistentes y estables cuando este se aplica en distintas ocasiones.

IV. **Constructo:** Elaboración teórica formulada para explicar un proceso social, psicológico o educativo.

V. **Correlación punto biserial:** Medida de consistencia que se utiliza en el análisis de reactivos, indica si hay una correlación entre el resultado de un reactivo con el resultado global del examen.

VI. **Criterio de evaluación:** Indicador de un valor aceptable sobre el cual se puede establecer o fundamentar un juicio de valor sobre el desempeño de una persona.

VII. **Desempeño:** Resultado obtenido por el sustentante en un proceso de evaluación o en un instrumento de evaluación educativa.

VIII. **Dificultad de un reactivo:** Indica la proporción de personas que responden correctamente el reactivo de un examen.

IX. **Distractores:** Opciones de respuesta incorrectas del reactivo de opción múltiple, que probablemente serán elegidas por los sujetos con menor dominio en lo que se evalúa.

X. **Dominio:** Conjunto de conocimientos, habilidades, destrezas, actitudes u otros atributos que tienen las siguientes propiedades: límites, extensión y definición. También se puede aplicar a contenidos, procedimientos u objetos.

XI. **Educación básica:** Tipo de educación que comprende los niveles de preescolar, primaria y secundaria en todas sus modalidades, incluyendo la educación indígena, la especial y la que se imparte en los centros de educación básica para adultos.

XII. **Educación media superior:** Tipo de educación que comprende el nivel de bachillerato, los demás niveles equivalentes a éste, así como la educación profesional que no requiere bachillerato o sus equivalentes.

XIII. **Equiparación:** Método estadístico que se utiliza para ajustar las puntuaciones de las formas o versiones de un mismo instrumento, de manera tal que al sustentante le sea indistinto, en términos de la puntuación que se le asigne, responder una forma u otra.

XIV. **Error estándar de medida:** Es la estimación de mediciones repetidas de un atributo evaluado sobre una misma persona a partir de un mismo instrumento y que tienden a distribuirse alrededor de un puntaje verdadero, el cual siempre es desconocido porque ninguna medida lo representa a la perfección.

XV. **Escala:** Conjunto de números, puntuaciones o medidas que pueden ser asignados a objetos o sucesos con propiedades específicas a partir de reglas definidas.

XVI. **Escalamiento:** Proceso a través del cual se construye una escala que facilita la interpretación de los resultados que se obtienen en uno o varios instrumentos de evaluación, colocando las puntuaciones de los distintos instrumentos o formas a una escala común.

XVII. **Especificaciones de tareas evaluativas o de reactivos:** Descripción detallada de las tareas específicas susceptibles de medición, que deben realizar las personas que contestan el instrumento de evaluación. Deben estar alineadas al constructo definido en el marco conceptual.

XVIII. **Estándar:** Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación. Constituyen el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XIX. **Estructura**. Está compuesta por los contenidos o aspectos disciplinares que mide un instrumento, así como el número y porcentaje relativo de reactivos o tareas evaluativas con que se integra el instrumento y su distribución.

XX. **Evaluación:** Proceso sistemático mediante el cual se recopila y analiza información, cuantitativa o cualitativa, sobre un objeto, sujeto o evento, con el fin de emitir juicios de valor al comparar los resultados con un referente previamente establecido. La información resultante puede ser empleada como insumo para orientar la toma de decisiones.

XXI. **Examen:** Instrumento de evaluación que se emplea para identificar el nivel de dominio de los sustentantes sobre un constructo específico.

XXII. **Formas de un instrumento**: Dos o más versiones de un instrumento que se consideran equivalentes, pues se construyen con los mismos contenidos y especificaciones estadísticas.

XXIII. **Instrumento de evaluación:** Herramienta de recolección de datos que suele tener distintos formatos, atendiendo a la naturaleza de la evaluación, por ejemplo, instrumentos de selección de respuesta, instrumentos de respuesta construida, cuestionarios, observaciones, portafolios, entre otros.

XXIV. **Instrumento de evaluación referido a un criterio:** Instrumento que permite comparar el desempeño de las personas evaluadas, con un estándar preestablecido.

XXV. **Jueceo:** Método en el cual se utiliza la opinión de expertos (denominados jueces) para valorar y calificar distintos aspectos, tales como las respuestas y ejecuciones de las personas que participan en una evaluación o la calidad de los reactivos, las tareas evaluativas y estándares de un instrumento.

XXVI. **Lista de prelación:** Orden descendente en que se enlistan los sustentantes con base en los resultados obtenidos en el proceso de evaluación**.**

XXVII. **Medición:** Proceso de asignación de valores numéricos a atributos de las personas, características de objetos o eventos de acuerdo con reglas específicas que permitan que sus propiedades puedan ser representadas cuantitativamente.

XXVIII. **Muestra:** Subconjunto de la población de interés que refleja las variables medidas en una distribución semejante a la de la población.

XXIX. **Multi-reactivo:** Conjunto de reactivos de opción múltiple que están vinculados a un planteamiento general, por lo que este último es indispensable para poder resolverlos.

XXX. **Nivel de desempeño:** Criterio conceptual que delimita el marco interpretativo de las puntuaciones obtenidas en una prueba y que refiere a lo que el sustentante es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.

XXXI. **Objeto de medida:** Conjunto de características o atributos que se miden en el instrumento de evaluación.

XXXII. **Parámetro estadístico:** Número que resume un conjunto de datos que se derivan del análisis de una cualidad o característica del objeto de estudio.

XXXIII. **Perfil:** Conjunto de características, requisitos, cualidades o aptitudes que deberá tener el sustentante a desempeñar un puesto o función descrito específicamente.

XXXIV. **Porcentaje de acuerdos inter-jueces:** Medida del grado en que dos jueces coinciden en la puntuación asignada a un producto o proyecto definido para evaluar, a través de una rúbrica, el desempeño de una persona.

XXXV. **Porcentaje de acuerdos intra-jueces:** Medida del grado en que el mismo juez, a través de dos o más mediciones repetidas a los mismos productos o proyectos evaluados con la misma rúbrica, coincide en las puntuaciones que les asignó.

XXXVI. **Protocolo**: Conjunto de reglas o normas que establecen cómo se deben realizar ciertas actividades o pasos, además de incorporar de manera detallada los procedimientos y los estándares que se deben cumplir. Para su realización es fundamental considerar las características especiales de cada evaluación.

XXXVII. **Punto de corte:** En instrumentos de evaluación con referencia a un estándar de desempeño, es la puntuación mínima o el criterio por alcanzar o a superar para considerar que el nivel de desempeño de una persona cumple con lo esperado y distinguirlo de otro que no.

XXXVIII. **Puntuación:** Valor numérico obtenido durante el proceso de medición.

XXXIX. **Reactivo:** Unidad básica de medida de un instrumento de evaluación que consiste en una pregunta o instrucción que requiere una respuesta del sujeto.

XL. **Rúbrica:** Herramienta que integra los criterios de ejecución a partir de los cuales se califica una tarea evaluativa.

XLI. **Sesgo:** Error en la medición de un atributo debido a una variable no controlada, como las diferencias culturales o lingüísticas de las personas evaluadas.

XLII. **Tarea evaluativa:** Unidad básica de medida de un instrumento de evaluación que consiste en la respuesta que construye una persona o en la ejecución de una actividad, que es susceptible de ser observada y graduada en su nivel de cumplimiento.

XLIII. **Teoría Clásica de los Tests (TCT)**: Teoría psicométrica que parte del supuesto de que el puntaje observado de una persona que responde un instrumento es la suma de su puntaje verdadero y un error aleatorio independiente del puntaje.

XLIV. **Validez:** Juicio valorativo integrador sobre el grado en que los fundamentos teóricos y las evidencias empíricas apoyan la interpretación de las puntuaciones de los instrumentos de evaluación.

1. Criterios técnicos para el análisis e integración de los instrumentos de evaluación

Uno de los aspectos fundamentales que debe llevarse a cabo antes de emitir cualquier resultado de un proceso de evaluación es el análisis psicométrico de los instrumentos que integran la evaluación, con el objetivo de verificar que cuentan con la calidad técnica necesaria para proporcionar resultados confiables, acordes con el objetivo de la evaluación.

Las técnicas empleadas para el análisis de un instrumento dependen de su naturaleza, de los objetivos específicos para los cuales fue diseñado, así como del tamaño de la población evaluada. Sin embargo, en todos los casos debe aportarse información sobre la dificultad y discriminación de sus reactivos o tareas evaluativas, así como la precisión del instrumento, los indicadores de consistencia interna o estabilidad del instrumento, los cuales, además de los elementos asociados a la conceptualización del objeto de medida, forman parte de las evidencias que servirán para valorar la validez de la interpretación de sus resultados. Estos elementos, deberán reportarse en el informe o manual técnico del instrumento.

Con base en los resultados de estos procesos de análisis deben identificarse las tareas evaluativas o los reactivos que cumplen con los criterios psicométricos especificados en este documento para integrar el instrumento, para calificar el desempeño de las personas evaluadas, con la mayor precisión posible.

Para llevar a cabo el análisis de los instrumentos de medición utilizados en el proceso de evaluación, es necesario que los distintos grupos de sustentantes de las entidades federativas queden equitativamente representados, dado que la cantidad de sustentantes por tipo de evaluación en cada entidad federativa es notoriamente diferente. Para ello, se definirá una muestra de sustentantes por cada instrumento de evaluación que servirá para analizar el comportamiento estadístico de los instrumentos y orientar los procedimientos descritos más adelante, y que son previos para la calificación. Para conformar dicha muestra, cada entidad federativa contribuirá con 500 sustentantes como máximo, y deberán ser elegidos aleatoriamente. Si hay menos de 500 sustentantes, todos se incluirán en la muestra (OECD; 2002, 2005, 2009, 2014). Si no se realizara este procedimiento, las decisiones sobre los instrumentos de evaluación, la identificación de los puntos de corte y los estándares de desempeño, se verían fuertemente influenciados, indebidamente, por el desempeño mostrado por aquellas entidades que se caracterizan por tener un mayor número de sustentantes.

Sobre la conformación de los instrumentos de evaluación

Con la finalidad de obtener puntuaciones de los aspirantes con el nivel de precisión requerido para los propósitos de los Concursos, los instrumentos de evaluación deberán tener las siguientes características:

Exámenes con reactivos de opción múltiple:

Los instrumentos de evaluación de carácter nacional deberán tener, al menos, 80 reactivos efectivos para calificación; en tanto que, los instrumentos complementarios, que atienden necesidades específicas de las entidades estatales, deberán tener una longitud igual o mayor a 60 reactivos efectivos para calificar. Todos los instrumentos deberán estar organizados jerárquicamente en tres niveles de desagregación: áreas, subáreas y temas, en donde:

- Las áreas deberán contar con al menos dos subáreas y, cada una de ellas, deberá tener al menos 20 reactivos efectivos para calificar.
- Las subáreas deberán considerar al menos dos temas, y cada uno de ellos deberá tener, al menos, 10 reactivos efectivos para calificar.
- Los temas deberán contemplar al menos dos contenidos específicos, los cuales estarán definidos en términos de especificaciones de reactivos. Cada especificación deberá ser evaluada al menos por un reactivo.

Exámenes de respuesta construida:

- Deberán estar organizados en, al menos, dos niveles de desagregación (áreas y subáreas; si fuera el caso, temas); el primero deberá contar, al menos, con dos conjuntos de contenidos específicos a evaluar.
- A partir del segundo nivel (o tercer nivel, si fuera el caso) de desagregación, se deberá contar con las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional.
- En las rúbricas o guías de calificación los distintos niveles o categorías de ejecución que se consignen deberán ser claramente distinguibles entre sí y con un diseño ordinal ascendente (de menor a mayor valor).

Criterios y parámetros estadísticos

Los instrumentos empleados para los Concursos de oposición para el Ingreso en EB y EMS para el ciclo escolar 2019-2020, deberán atender los siguientes criterios (Cook y Beckman 2006; Downing, 2004; Stemler y Tsai, 2008) con, al menos, los valores de los parámetros estadísticos indicados a continuación:

En el caso de los instrumentos de evaluación basados en reactivos de opción múltiple:

- La respuesta correcta deberá tener una dificultad clásica de 10% a 90% y una correlación punto biserial corregida igual o mayor que 0.15.
- Los distractores deberán tener correlaciones punto biserial negativas.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.80.

Para los instrumentos con menos de 100 sustentantes, la selección de los reactivos con los cuales se va a calificar, se debe llevar a cabo con base en el siguiente procedimiento: cada reactivo tiene que ser revisado por, al menos, tres jueces: dos expertos en contenido y un revisor técnico, considerando los siguientes aspectos: *calidad del contenido del reactivo, adecuada construcción técnica, correcta redacción y atractiva presentación de lo que se evalúa*.

En todos los casos en los que sea factible estimar los parámetros estadísticos de los reactivos, esta información debe proporcionarse a los jueces con el objetivo de que les permita fundamentar sus decisiones y ejercer su mejor juicio profesional.

En el caso de los instrumentos basados en tareas evaluativas o en reactivos de respuesta construida y que serán calificados con rúbrica:

- La correlación corregida entre cada aspecto evaluado con la puntuación global deberá ser igual o mayor que 0.20.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.80.

Considerando las decisiones de los jueces que calificaron los instrumentos de respuesta construida a través de rúbrica se debe atender lo siguiente:

- El porcentaje de acuerdos inter-jueces deberá ser igual o mayor que 60%.
- El porcentaje de acuerdos intra-jueces deberá ser igual o mayor que 60% considerando, al menos, cinco medidas repetidas seleccionadas al azar, es decir, para cada juez se deben seleccionar al azar cinco sustentantes, a quienes el juez debe calificar en dos ocasiones a través de la estrategia "doble ciego", a fin de que, en su segunda calificación, el juez no identifique el producto del sustentante. Estas mediciones deberán aportarse antes de emitir la calificación definitiva de los sustentantes, a fin de salvaguardar la confiabilidad de la decisión.

Si se diera el caso de que en algún instrumento no se cumpliera con los criterios y parámetros estadísticos antes indicados, la Junta de Gobierno del INEE determinará lo que procede, buscando salvaguardar el constructo del instrumento que fue aprobado por el Consejo Técnico y atendiendo al marco jurídico aplicable.

2. Procedimiento para el establecimiento de puntos de corte y estándares de desempeño

Un paso crucial en el desarrollo y uso de los instrumentos de evaluación de naturaleza criterial, como es el caso de los que se utilizan para evaluar a los sustentantes de los Concursos de oposición para el ingreso al Servicio Profesional Docente en Educación Básica y Educación Media Superior, es el establecimiento de los puntos de corte que dividen el rango de calificaciones para diferenciar entre niveles de desempeño.

En los instrumentos de evaluación de tipo criterial, la calificación obtenida por cada sustentante se contrasta con un estándar de desempeño establecido por un grupo de expertos que describe el nivel de competencia requerido para algún propósito determinado, es decir, los conocimientos y habilidades que, para cada instrumento de evaluación, se consideran indispensables para un desempeño adecuado en la función profesional. En este sentido el estándar de desempeño delimita el marco interpretativo de las puntuaciones obtenidas en un instrumento por los sustentantes.

El procedimiento para el establecimiento de puntos de corte y estándares de desempeño incluye tres fases, las cuales se describen a continuación:

Primera fase

Con el fin de contar con un marco de referencia común para los distintos instrumentos de evaluación, se deberán establecer descriptores genéricos de los niveles de desempeño que se utilizarán y **cuya única función** es orientar a los comités académicos en el trabajo del desarrollo de los descriptores específicos de cada instrumento, tales que les permita a los docentes tener claros elementos de retroalimentación para conocer sus fortalezas y áreas de oportunidad identificadas a partir de los resultados de cada instrumento sustentado.

Para todos los instrumentos se utilizarán tres niveles de desempeño posibles: Nivel I (NI), Nivel II (NII) y Nivel III (NIII). Los descriptores genéricos para cada uno de ellos se indican en la Tabla 1.

Tabla 1. Descriptores genéricos de los niveles de desempeño

Nivel de desempeño	Descriptor
Nivel I (NI)	Dominio insuficiente de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño docente.
Nivel II (NII)	Dominio suficiente y organizado de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño docente.
Nivel III (NIII)	Dominio suficiente y organizado de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño docente, con *amplia capacidad de utilizarlos en una diversidad de situaciones didácticas.*

Segunda fase

En esta fase se establecerán los puntos de corte y deberán participar los comités académicos específicos para el instrumento de evaluación que se esté trabajando. Dichos comités se deberán conformar, en su conjunto, con especialistas que han participado en el diseño de los instrumentos y cuya pluralidad sea representativa de la diversidad cultural en que se desenvuelve la acción educativa del país. En todos los casos, sus miembros deberán ser capacitados específicamente para ejercer su mejor juicio profesional a fin de identificar cuál es la puntuación requerida para que el sustentante alcance un determinado nivel o estándar de desempeño.

Los insumos que tendrán como referentes para el desarrollo de esta actividad serán, la documentación que describe la estructura de los instrumentos, las especificaciones, los ejemplos de tareas evaluativas o de reactivos incluidos en las mismas y las rúbricas utilizadas para la calificación. En todos los casos, los puntos de corte se referirán a la ejecución típica o esperable de un sustentante hipotético, con un desempeño mínimamente aceptable, para cada uno de los niveles. Para ello, se deberá determinar, para cada tarea evaluativa o reactivo considerado en el instrumento, cuál es la probabilidad de que dicho sustentante hipotético lo responda correctamente y, con base en la suma de estas probabilidades, establecer la calificación mínima requerida o punto de corte, para cada nivel de desempeño (Angoff, 1971).

Una vez establecidos los puntos de corte que dividen el rango de calificaciones para diferenciar los niveles de desempeño en cada instrumento, se deberán describir los conocimientos y las habilidades específicos que están implicados en cada nivel de desempeño, es decir, lo que dicho sustentante conoce y es capaz de hacer.

Tercera fase

En la tercera fase se llevará a cabo un ejercicio de retroalimentación a los miembros de los comités académicos con el fin de contrastar sus expectativas sobre el desempeño de la población evaluada, con la distribución de sustentantes que se obtiene en cada nivel de desempeño al utilizar los puntos de corte definidos en la segunda fase, a fin de determinar si es necesario realizar algún ajuste en la decisión tomada con anterioridad y, de ser el caso, llevar a cabo el ajuste correspondiente.

Los jueces deberán estimar la tasa de sustentantes que se esperaría en cada nivel de desempeño y comparar esta expectativa con los datos reales de los sustentantes, una vez aplicados los instrumentos. Si las expectativas y los resultados difieren a juicio de los expertos, deberá definirse un punto de concordancia para la determinación definitiva del punto de corte asociado a cada nivel de desempeño en cada uno de los instrumentos, siguiendo el método propuesto por Beuk (1984).

Esta tercera fase se llevará a cabo solamente para aquellos instrumentos de evaluación en los que el tamaño de la población evaluada sea igual o mayor a 100 sustentantes. Si la población es menor a 100 sustentantes, los puntos de corte serán definidos de acuerdo con lo descrito en la segunda fase.

Si se diera el caso de que algún instrumento no cumpliera con el criterio de confiabilidad indicado en el apartado previo, la Junta de Gobierno del INEE determinará el procedimiento a seguir para el establecimiento de los puntos de corte correspondientes, atendiendo al marco jurídico aplicable.

3. Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos del proceso de evaluación

Todos los sustentantes que participen en los Concursos de oposición para el Ingreso al SPD en Educación Básica y Educación Media Superior 2019-2020, deberán sustentar, al menos, dos exámenes. Cada sustentante recibirá los resultados de cada uno de los exámenes que haya presentado, así como el resultado integrado de todo el proceso de evaluación.

Una vez que se han establecido los puntos de corte en cada examen, el resultado del sustentante será ubicado en uno de los tres niveles de desempeño en función de la puntuación alcanzada. Esto implica que su resultado será comparado contra el estándar previamente establecido, con independencia de los resultados obtenidos por el conjunto de sustentantes que presentaron el examen.

3.1 Proceso para la equiparación de instrumentos de evaluación

Cuando el proceso de evaluación implica la aplicación de un instrumento en diversas ocasiones en un determinado periodo, en especial si sus resultados tienen un alto impacto, es indispensable el desarrollo y uso de formas o versiones del instrumento que sean equivalentes a fin de garantizar que, independientemente del momento en que un aspirante participe en el proceso de evaluación, no tenga ventajas o desventajas de la forma o versión que responda. Por esta razón, es necesario un procedimiento que permita hacer equivalentes los resultados obtenidos en diferentes formas o versiones de un mismo instrumento.

Para que dos formas de un instrumento de evaluación puedan ser equiparadas, se deben cubrir los siguientes requerimientos:

- Compartir las mismas características técnicas: estructura, especificaciones de reactivos, número de reactivos (longitud del instrumento) y un subconjunto de reactivos comunes (reactivos ancla), que en cantidad no deberá ser menor al 30% ni mayor al 50% de la totalidad de reactivos efectivos para calificar.
- Contar con una confiabilidad semejante.
- Los reactivos que constituyen el ancla deberán ubicarse en la misma posición relativa dentro de cada forma, y deberán quedar distribuidos a lo largo de todo el instrumento.
- La modalidad en la que se administren las formas deberá ser la misma para todos los aspirantes (por ejemplo, en lápiz y papel o en computadora).

Si el número de sustentantes es de al menos 100 en las distintas formas en que se llevará a cabo la equiparación, se utilizará el método de equiparación lineal para puntajes observados. Si el número de sustentantes es menor de 100 en alguna de las formas, se utilizará el método de equiparación de identidad (ver anexo técnico).

3.2 Escala utilizada para reportar los resultados

En cada plan de evaluación es indispensable definir la escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. En las escalas referidas a criterio, cada calificación en la escala representa un nivel particular de desempeño referido a un estándar previamente definido en un campo de conocimiento o habilidad específico.

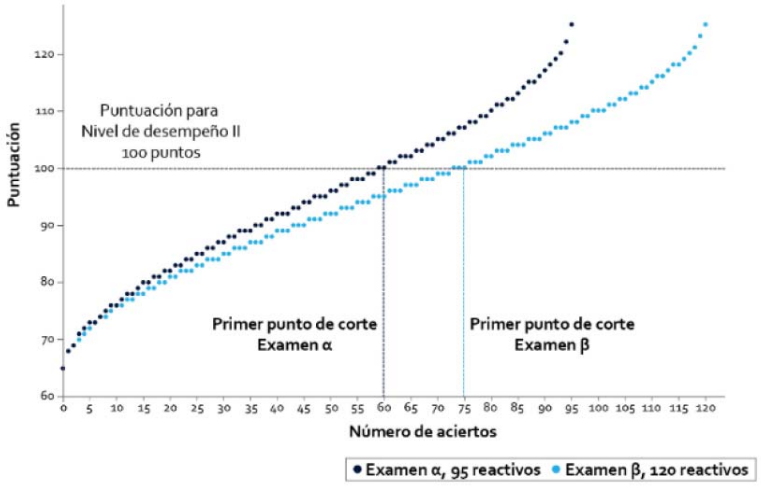
El escalamiento que se llevará a cabo en este proceso de evaluación, permitirá construir una métrica común para todos los instrumentos que se administrarán. Consta de dos transformaciones, la primera denominada doble arcoseno, que permite estabilizar la magnitud de la precisión de las puntuaciones a lo largo de la escala; la segunda transformación es lineal y ubica el punto de corte del nivel de desempeño II en un mismo valor para todos los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos[1]).

---

[1] Pueden encontrarse ligeras variaciones en este rango debido a que la escala es aplicable a múltiples instrumentos con características diversas, tales como la longitud, el tipo de instrumento y su nivel de precisión, diferencias entre los puntos de corte que atienden a las particularidades de los contenidos que se evalúan, entre otras. Por otra parte, para realizar el escalamiento, el sustentante debe, al menos, alcanzar un acierto en el examen; en caso contrario, se reportará como cero puntos y obtendrá NI.
En el caso de que el sustentante no presente alguno de los instrumentos de la evaluación, su resultado en ese instrumento será "NP: no presentó" y únicamente tendrá la devolución en aquellos instrumentos en los que haya participado y de los que se cuente con información, su resultado en el proceso de evaluación será "No idóneo" y no formará parte de los grupos de desempeño ni de las listas de prelación que publica la CNSPD a través de su página web.
Cuando el sustentante no presente ninguno de los instrumentos de la evaluación, su resultado en cada instrumento será "NP: no presentó", su resultado en el proceso de evaluación será "No se presentó a la evaluación" y no formará parte de los grupos de desempeño ni de las listas de prelación que publica la CNSPD a través de su página web.
Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

Al utilizar esta escala, diferente a las escalas que se utilizan para reportar resultados de aprendizaje en el aula (de 5 a 10 o de 0% a 100%, donde el 6 o 60% de aciertos es aprobatorio), se evita que se realicen interpretaciones equivocadas de los resultados obtenidos en los exámenes, en virtud de que en los exámenes del SPD cada calificación representa un nivel particular de desempeño respecto a un estándar previamente definido, el cual puede implicar un número de aciertos diferente en cada caso.

En la siguiente gráfica puede observarse el número de aciertos obtenido en dos instrumentos de longitudes diferentes y con puntos de corte distintos que, a partir del escalamiento, es posible graficar en una misma escala, trasladando el primer punto de corte a 100 puntos, aun cuando en cada examen el punto de corte refiera a número de aciertos diferente. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



4. Resultado del proceso de evaluación: conformación de los grupos de desempeño

Dado que en cada instrumento se miden dominios diferentes y se atiende una lógica propia de diseño, construcción e incluso calificación, en ningún caso podrán sumarse el número de aciertos de cada examen para generar una puntuación global de todo el proceso de evaluación. Por esta razón, para determinar la idoneidad del resultado de cada aspirante se deberán integrar los resultados de todos los instrumentos sustentados, bajo el criterio de que:

> El sustentante con un resultado Idóneo en el proceso de evaluación será aquel que obtenga, al menos, el nivel de desempeño II (NII) en todos y cada uno de los instrumentos de evaluación que lo constituyen, según se define en los lineamientos del Concurso.

Cada sustentante conocerá su resultado integrado de todo el proceso de evaluación, así como sus resultados en cada uno de los exámenes que haya presentado.

Conformación de los grupos de desempeño

Con el conjunto de sustentantes que obtengan un resultado Idóneo en el proceso de evaluación, se conformarán grupos de desempeño en función de la combinación de resultados alcanzados del nivel de desempeño II o III (NII o NIII) en los instrumentos considerados en dicho proceso de evaluación. Los grupos de desempeño son el primer criterio de ordenamiento para la integración de las listas de prelación con los resultados idóneos que publica la CNSPD a través de su página web.

Como ejemplo, los grupos de desempeño en los procesos de evaluación que constan de dos instrumentos, se conformarán de la manera siguiente:

> El **primer grupo de desempeño (A)**, se conformará con aquellos aspirantes que alcancen el nivel de desempeño III (NIII) en los dos exámenes involucrados en su proceso de evaluación.
>
> El **segundo grupo de desempeño (B)**, se conformará por los aspirantes que alcancen el nivel
> de desempeño II (NII) en uno de los exámenes y el nivel de desempeño III (NIII) en el otro examen.
>
> El **tercer grupo de desempeño (C)**, se conformará por los aspirantes que alcancen el nivel de desempeño II (NII) en los dos exámenes.

Para los casos en los que el proceso de evaluación considere tres instrumentos, los grupos de desempeño se deberán integrar con la misma lógica de ordenamiento en función del nivel de desempeño alcanzado por los sustentantes en cada uno de los instrumentos. A continuación, se presentan las tablas 2a y 2b, indicando los grupos de desempeño que se organizan en función del número de exámenes y los niveles de desempeño II y III obtenidos en cada uno de ellos.

Tabla 2a. Grupos de desempeño con dos exámenes

Criterios para formar parte de un grupo de desempeño		Educación básica
Grupos de desempeño	**Descripción**	Presentan dos exámenes los aspirantes a plazas docentes y técnico docentes: 1) Examen nacional de conocimientos y habilidades para la práctica docente y 2) Examen nacional de habilidades intelectuales y responsabilidades ético-profesionales.
A	En los dos exámenes obtuvo el nivel de desempeño NIII	**Educación media superior**
B	En un examen obtuvo el nivel de desempeño NIII y en el otro el nivel de desempeño NII	Presentan dos exámenes los aspirantes a plazas docentes y técnico docentes de las disciplinas asociadas al componente profesional técnico: 1) Examen de habilidades docentes y 2) Plan de clase.
C	En los dos exámenes obtuvo el nivel de desempeño NII	

Tabla 2b. Grupos de desempeño con tres exámenes

Criterios para formar parte de un grupo de desempeño		Educación básica
Grupos de desempeño	**Descripción**	Presentan tres exámenes los aspirantes a plazas docentes con requerimientos particulares de acuerdo con la especificación de la Entidad federativa correspondiente: 1) Examen nacional de conocimientos y habilidades para la práctica docente, 2) Examen nacional de habilidades intelectuales y responsabilidades ético-profesionales. y 3) Examen complementario o adicional, con base en lo que se indique en la convocatoria.
A	En los tres exámenes obtuvo el nivel de desempeño NIII	**Educación media superior**
B	En dos exámenes obtuvo el nivel de desempeño NIII y en el otro el nivel de desempeño NII	Presentan tres exámenes los aspirantes a plazas docentes del componente disciplinar. Los instrumentos son: 1) Examen de contenidos disciplinares, 2) Examen de habilidades docentes y 3) Plan de clase.
C	En un examen obtuvo el nivel de desempeño NIII y en los otros dos el nivel de desempeño NII	
D	En los tres exámenes obtuvo el nivel de desempeño NII	

5. Integración de las listas de prelación

Las listas de prelación que son publicadas por la CNSPD se integrarán únicamente con los sustentantes que obtengan un resultado Idóneo en su proceso de evaluación. La lista se ordenará, en primer término, considerando los grupos de desempeño, iniciando con el grupo A, después el B y así de manera sucesiva.

Posteriormente, al interior de cada grupo, se ordenará la lista considerando, primero, la puntuación obtenida por los sustentantes en el examen de mayor relevancia o jerarquía dentro del conjunto de instrumentos implicados en el proceso de evaluación, después la calificación obtenida en el instrumento que le sigue en relevancia, y así sucesivamente.

El último criterio de ordenación está dado por las puntuaciones obtenidas por los sustentantes en contenidos de segundo nivel (por ejemplo, las áreas) de mayor importancia del examen de mayor jerarquía, después el puntaje obtenido en el área que le sigue de importancia, y así sucesivamente (la mayor relevancia es el 1, le sigue el 2, etc.).

De manera excepcional, si después de llevar a cabo el proceso anterior para generar las listas de prelación se observan empates entre los resultados de algunos sustentantes, se recurrirá a la misma lógica de ordenamiento (ahora con conteo de aciertos), tomando como referente el segundo nivel de desagregación de los contenidos específicos, que cuenten con la mayor cantidad de reactivos y que formen parte del examen que ha sido considerado como el de mayor relevancia. La jerarquía estará dada por el orden de la estructura del instrumento.

A continuación, se describe la jerarquía de los instrumentos y la relevancia de los contenidos específicos en cada uno de ellos para Educación Básica y Educación Media Superior, referidos en los párrafos anteriores.

Criterios específicos de ordenamiento para EB

Jerarquía de los instrumentos de evaluación para los docentes

Los grupos de desempeño se definen de acuerdo con lo estipulado en la Tabla 2a, toda vez que el proceso de evaluación considera dos exámenes:

1º Examen nacional de conocimientos y habilidades para la práctica docente

2º Examen nacional de habilidades intelectuales y responsabilidades ético-profesionales

La jerarquización de los contenidos específicos de primer nivel en cada uno de los instrumentos de evaluación se presenta en la Tabla 3.

Tabla 3. Jerarquía de los contenidos específicos de los instrumentos de evaluación para Docentes en Educación Básica, en donde el orden de mayor a menor relevancia es 1, 2, 3, 4 y 5

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Preescolar*		Primaria		Secundaria			Educación especial	Educación física	Inglés
		General	Indígena	General	Indígena	General	Artes	Telesecundaria**			
Examen Nacional de Conocimientos y Habilidades para la Práctica Docente	Intervención didáctica	1°	1°	2°	2°	2°	2°	1°	2°	2°	2°
	Aspectos del currículo	2°	2°	1°	1°	1°	1°	2°	1°	1°	1°
Examen Nacional de Habilidades Intelectuales y Responsabilidades Ético Profesionales	Compromiso Ético	3°	3°	3°	3°	3°	3°	3°	3°	3°	3°
	Mejora profesional	4°	4°	4°	4°	4°	4°	4°	4°	4°	4°
	Gestión escolar y vinculación con la comunidad	5°	5°	5°	5°	5°	5°	5°	5°	5°	5°

*Intervención didáctica en la educación preescolar -en las modalidades de general e indígena- es prioritaria porque se requiere considerar el desarrollo de los alumnos para incidir en la manera en que se construyen los aprendizajes, respetando sus procesos cognitivos. De igual modo, porque los niños están en un periodo en el que la relación afectiva determina su seguridad, desenvolvimiento y mantiene su curiosidad por seguir aprendiendo. En este sentido, las habilidades para la práctica docente son fundamentales para la educación en este nivel educativo. Por lo tanto, siguiendo el orden de mayor relevancia, la prioridad 1 es la Intervención didáctica y la 2 es Aspectos del currículo.

**En la modalidad de Telesecundaria, el docente imparte todas las asignaturas, y el contenido curricular está determinado por los programas de televisión y los programas impresos. Por lo tanto, siguiendo el orden de mayor relevancia, la primera prioridad en este caso es Intervención didáctica, la cual es necesaria para que el docente articule el contenido curricular de las diferentes asignaturas con un sentido formativo general, y la segunda prioridad es Aspectos del currículo.

Jerarquía de los instrumentos de evaluación para los técnicos docentes

Se definen los grupos de desempeño de acuerdo con la Tabla 2a, ya que el proceso de evaluación considera dos exámenes:

1º Examen nacional de conocimientos y habilidades para la práctica docente

2º Examen nacional de habilidades intelectuales y responsabilidades ético-profesionales

La jerarquización de los contenidos específicos de primer nivel de los instrumentos de evaluación se presenta en la Tabla 4.

Tabla 4. Jerarquía de los contenidos específicos de los instrumentos de evaluación para técnicos docentes en Educación Básica, en donde el orden de mayor a menor relevancia es 1, 2, 3, 4 y 5.

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Maestro de taller de lectura y escritura. Preescolar, Primaria y Secundaria	Maestro de enseñanza artística. Primaria	Maestro de taller. Primaria	Promotor de TIC. Primaria	Maestro de aula de medios. Secundaria
Examen Nacional de Conocimientos y Habilidades para la Práctica Docente	Intervención didáctica	2°	2°	2°	2°	2°
	Aspectos del currículo	1°	1°	1°	1°	1°
Examen Nacional de Habilidades Intelectuales y Responsabilidades Ético Profesionales	Compromiso Ético	3°	3°	3°	3°	3°
	Mejora profesional	4°	4°	4°	4°	4°
	Gestión escolar y vinculación con la comunidad	5°	5°	5°	5°	5°

Jerarquía de los instrumentos de evaluación para docentes con una evaluación complementaria

Para los casos en que el proceso de evaluación considere un instrumento complementario, los grupos de desempeño se definen con base en lo que se indica en la Tabla 2b, ya que se consideran tres exámenes:

1º Examen nacional de conocimientos y habilidades para la práctica docente

2º Examen nacional de habilidades intelectuales y responsabilidades ético-profesionales

3º Examen complementario

La jerarquización de los contenidos específicos de segundo nivel de los instrumentos de evaluación contemplará únicamente a los exámenes nacionales, en el orden descrito previamente en la Tabla 3.

Criterios específicos de ordenamiento para EMS

Jerarquía de los instrumentos de evaluación para docentes y técnicos docentes de las disciplinas asociadas al componente profesional técnico.

Los grupos de desempeño se definen de acuerdo con lo estipulado en la Tabla 2a, ya que se consideran 2 exámenes:

1º Examen de habilidades docentes

2º Plan de clase

La jerarquización de los contenidos específicos de segundo nivel de los instrumentos de evaluación considerará como criterio de relevancia el orden secuencial en que se organizan en las estructuras de los instrumentos de evaluación. Esto aplicará para el Examen de habilidades docentes, como se describe en la Tabla 5.

Tabla 5. Jerarquía de los contenidos específicos de evaluación para docentes y técnicos docentes de las disciplinas asociadas al componente profesional técnico en EMS, en donde el orden de mayor a menor relevancia es 1 y 2.

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Jerarquía
Examen de habilidades docentes	**Organización de la práctica docente**	1°
	Intervención docente	2°

El instrumento de evaluación **Plan de clase** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

Jerarquía de los instrumentos de evaluación para docentes del componente disciplinar

Los grupos de desempeño se definen de acuerdo con lo estipulado en la Tabla 2b, toda vez que el proceso de evaluación considera tres exámenes:

1º Examen de contenidos disciplinares

2º Examen de habilidades docentes

3º Plan de clase

La jerarquización de los contenidos específicos de segundo nivel de los instrumentos de evaluación considerará como criterio de relevancia el orden secuencial en que se organizan en las estructuras de los instrumentos de evaluación. Esto aplicará para el Examen de contenidos disciplinares, según su disciplina (Tabla 6), y el Examen de habilidades docentes, en el orden descrito previamente en la Tabla 5.

Tabla 6. Jerarquía de los contenidos específicos de evaluación para docentes del componente disciplinar en EMS, en donde el orden de mayor a menor relevancia es 1, 2 y 3.

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Jerarquía
Examen de contenidos disciplinares Administración	**Bases de la administración**	1°
	Actividades empresariales	2°
	Habilidades del emprendedor	3°
Examen de contenidos disciplinares Biología	**Estructura y dinámica de la vida**	1°
	Ciencia, biotecnología y sociedad	2°
Examen de contenidos disciplinares Derecho	**Clasificación, jerarquía y principios del derecho**	1°
	Clasificación de las normas y fuentes del derecho	2°
	Teorías del estado	3°
Examen de contenidos disciplinares Dibujo	**Fundamentos de la representación gráfica**	1°
	Sistemas de representación	2°
Examen de contenidos disciplinares Economía	**Antecedentes económicos**	1°
	Sistema económico	2°
Examen de contenidos disciplinares Etimologías grecolatinas	**Etimologías griegas del español**	1°
	Etimologías latinas del español	2°
	Aplicación de las etimologías	3°
Examen de contenidos disciplinares Física	**Mecánica**	1°
	Fluidos y calor	2°
	Electromagnetismo y ondas	3°
Examen de contenidos disciplinares Geografía	**Dinámica natural del espacio geográfico**	1°
	Dinámica social del espacio geográfico	2°
Examen de contenidos disciplinares Historia	**Marco teórico de la historia. México prehispánico y la conquista de México**	1°
	Desarrollo histórico de México y del mundo	2°
Examen de contenidos disciplinares Historia del arte	**Interpretación del arte**	1°
	La evolución del arte	2°
Examen de contenidos disciplinares Humanidades	**Metodología del pensamiento**	1°
	Disciplinas filosóficas	2°
Examen de contenidos disciplinares Informática	**Fundamentos de computación e internet**	1°
	Ingeniería de software y redes de cómputo	2°
Examen de contenidos disciplinares Lectura y expresión escrita	**Lectura**	1°
	Escritura	2°
	Comunicación	3°
Examen de contenidos disciplinares Matemáticas	**Algebra, geometría y trigonometría**	1°
	Geometría analítica, cálculo y estadística	2°
Examen de contenidos disciplinares Química	**Materia y energía**	1°
	Reacciones químicas, fisicoquímica y biomoléculas	2°
Examen de contenidos disciplinares Sociología-política	**Conocimiento social**	1°
	Cambio y realidad social	2°

Examen de contenidos disciplinares Psicología	**Bases psicofisiológicas y teorías de la psicología**	1°
	Procesos psicosociales y áreas de aplicación e investigación en psicología	2°
Examen de contenidos disciplinares Literatura	**Textos literarios**	1°
	Historia de la literatura	2°
Examen de contenidos disciplinares Ciencias de la salud	**Salud**	1°
	Prevención	2°
Examen de contenidos disciplinares Metodología de la investigación	**Fundamentos de la investigación**	1°
	Diseño del proyecto de investigación	2°

El instrumento de evaluación **Plan de clase** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

Finalmente, cualquier situación no prevista en los presentes criterios técnicos será analizada por la Junta de Gobierno del INEE para emitir una determinación, según corresponda con el marco normativo vigente.

Sobre la integralidad de la evaluación para emitir la calificación

Dado que los presentes criterios técnicos se han definido con el objetivo de aportar evidencia para la validez de las inferencias que se desean obtener a partir de los datos recopilados, es fundamental señalar que, ***en ningún caso, se puede considerar solamente un instrumento para integrar el resultado de los sustentantes conforme al diseño de la evaluación***. Lo anterior debido a que la evidencia empírica que resulte del análisis psicométrico de los instrumentos debe mostrar que, una vez que éstos fueron aplicados, cumplen con la calidad técnica indispensable para emitir un resultado toda vez que la integración de los resultados de la evaluación debe permitir establecer inferencias válidas sobre el desempeño y competencias de los sustentantes evaluados.

Anexo técnico

El propósito de este anexo es detallar los aspectos técnicos específicos de los distintos procedimientos que se han enunciado en el cuerpo del documento, así como de brindar mayores elementos para su entendimiento y fundamento metodológico.

Protocolo de calificación por jueces a través de rúbricas

A continuación, se presenta un protocolo que recupera propuestas sistemáticas de la literatura especializada (Jonsson y Svingby, 2007; Rezaei y Lovorn, 2010; Stemler y Tsai, 2008; Stellmack, et. al, 2009).

1. Se reciben las evidencias de evaluación de los sustentantes, mismas que deben cumplir con las características solicitadas por la autoridad educativa.

2. Se da a conocer a los jueces la rúbrica de calificación y se les capacita para su uso.

3. Las evidencias de los sustentantes son asignadas de manera aleatoria a los jueces, por ejemplo se pueden considerar *redes no dirigidas*; intuitivamente, una red no dirigida puede pensarse como aquella en la que las conexiones entre los nodos siempre son simétricas (si A está conectado con B, entonces B está conectado con A y sucesivamente con los *n* número de jueces conectados entre sí), este tipo de asignación al azar permite contar con indicadores iniciales de cuando un juez está siendo reiteradamente "estricto" o reiteradamente "laxo" en la calificación, lo cual ayudará a saber si es necesario volver a capacitar a alguno de los jueces y permitirá obtener datos de consistencia inter-juez.

4. Cada juez califica de manera individual las evidencias sin conocer la identidad de los sustentantes o cualquier otro dato que pudiera alterar la imparcialidad de la decisión del juez.

5. Los jueces emiten la calificación de cada sustentante, seleccionando la categoría de ejecución que consideren debe recibir el sustentante para cada uno de los aspectos a evaluar que constituyen la rúbrica, esto en una escala ordinal (por ejemplo: de 0 a 3, de 0 a 4, de 1 a 6, etc.), lo pueden hacer en un formato impreso o electrónico a fin de conservar dichas evidencias.

6. Si existen discrepancias entre los dos jueces en cuanto a la asignación de categorías en algunos aspectos a evaluar se deben tomar decisiones al respecto. A continuación, se muestran orientaciones para esta toma de decisiones:

a. Cuando la discrepancia corresponde a categorías de ejecución contiguas (por ejemplo: 1-2) se asigna la categoría superior. Esto permite favorecer al sustentante ante dicho desacuerdo entre los jueces.

b. Cuando son categorías no contiguas de la rúbrica:

- Si existe solamente una categoría en medio de las decisiones de los jueces (por ejemplo: 1-3), se asigna al sustentante la categoría intermedia. No se deben promediar los valores asignados a las categorías.
- Si existe más de una categoría en medio de las decisiones de los jueces (por ejemplo: 1-4), se debe solicitar a los jueces que verifiquen si no hubo un error al momento de plasmar su decisión. En caso de no haber ajustes por este motivo, se requiere la intervención de un tercer juez, quien debe asignar la categoría de ejecución para cada uno de los aspectos a evaluar; la categoría definitiva que se asigna al sustentante en cada aspecto a evaluar debe considerar las decisiones de los dos jueces que den mayor puntaje total al sustentante, si existe discrepancia en algún aspecto a evaluar se asigna la categoría superior, a fin de favorecer al sustentante ante dicho desacuerdo entre los jueces.

7. Los jueces firman la evidencia con las asignaciones de categorías definitivas en cada aspecto a evaluar.

8. La calificación global del sustentante se determina de la siguiente forma:

a. Se identifica la categoría asignada al sustentante en cada aspecto a evaluar.

b. Se identifica el valor asignado a cada categoría de la rúbrica.

c. La suma de los valores es el resultado de la calificación.

9. Las asignaciones de categorías en cada aspecto a evaluar para emitir la calificación global definitiva del sustentante son plasmadas en algún formato impreso o electrónico, con la debida firma, autógrafa o electrónica de los jueces, a fin de que queden resguardadas como evidencia del acuerdo de la calificación definitiva del proceso de jueceo.

En el caso de los **instrumentos para lenguas indígenas**, que son calificados *in situ* por jueces, se calculará la puntuación final que otorga cada juez con base en sus asignaciones, la que resulte mayor de entre los dos jueces será la que se considere como puntuación definitiva del sustentante.

Métodos para establecer puntos de corte y niveles de desempeño

Método de Angoff

El método de Angoff está basado en los juicios de los expertos sobre los reactivos y contenidos que se evalúan a través de exámenes. De manera general, el método considera que el punto de corte se define a partir de la ejecución promedio de un sustentante hipotético que cuenta con los conocimientos, habilidades o destrezas que se consideran indispensables para la realización de una tarea en particular; los jueces estiman, para cada pregunta, cuál es la probabilidad de que dicho sustentante acierte o responda correctamente.

Procedimiento

Primero se juzgan algunas preguntas, con tiempo suficiente para explicar las razones de las respuestas al grupo de expertos y que les permite homologar criterios y familiarizarse con la metodología.

Posteriormente, se le solicita a cada juez que estime la probabilidad mínima de que un sustentante conteste correctamente un reactivo, el que le sigue y así hasta concluir con la totalidad de los reactivos, posteriormente se calcula el puntaje esperado (*raw score*: la suma de estas probabilidades multiplicadas por uno para el caso de reactivos –toda vez que cada reactivo vale un punto–; o bien, la suma de estas probabilidades multiplicadas por el valor máximo posible de las categorías de la rúbrica). Las decisiones de los jueces se promedian obteniendo el punto de corte. La decisión del conjunto de jueces pasa por una primera ronda para valorar sus puntos de vista en plenaria y puede modificarse la decisión hasta llegar a un acuerdo en común.

Método de Beuk

En 1981, Cess H. Beuk propuso un método para establecer estándares de desempeño, el cual busca equilibrar los juicios de expertos basados solamente en las características de los instrumentos de evaluación, lo que mide y su nivel de complejidad, con los juicios que surgen del análisis de resultados de los sustentantes una vez que un instrumento de evaluación es administrado.

Procedimiento

En el cuerpo del documento se señalaron tres fases para el establecimiento de los puntos de corte. Para completar la tercera fase, es necesario recolectar con antelación las respuestas a dos preguntas dirigidas a los integrantes de los distintos comités académicos especializados involucrados en el diseño de las evaluaciones y en otras fases del desarrollo del instrumento. Las dos preguntas son:

a) ¿Cuál es el mínimo nivel de conocimientos o habilidades que un sustentante debe tener para aprobar el instrumento de evaluación? (expresado como porcentaje de aciertos de todo el instrumento, k).

b) ¿Cuál es la tasa de aprobación de sustentantes que los jueces estiman que aprueben el instrumento? (expresado como porcentaje, v).

Para que los resultados de la metodología a implementar sean estables e integren diferentes enfoques que contribuyan a la diversidad cultural, se deberán recolectar las respuestas de, al menos, 30 especialistas integrantes de los diferentes comités académicos que hayan participado en el diseño y desarrollo de los instrumentos.

Adicionalmente, se debe contar con la distribución de los sustentantes para cada posible punto de corte, con la finalidad de hacer converger el juicio de los expertos con la evidencia empírica.

Los pasos por seguir son los siguientes:

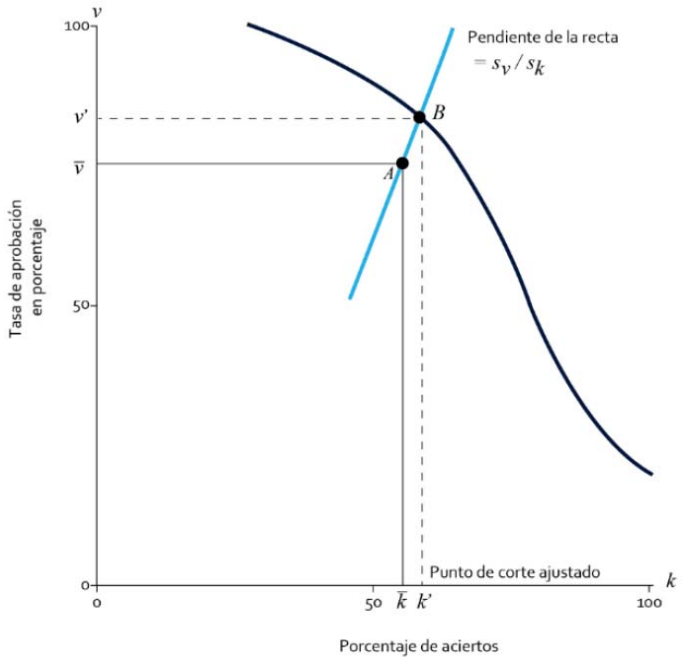
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas ($\bar{k}$, $\bar{v}$), (ver siguiente figura).

2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

3. Se calcula la desviación estándar de k y v (s_k y s_v).

4. A partir del punto A se proyecta una recta con pendiente s_v/s_k hasta la curva de distribución empírica (del paso 2). El punto de intersección entre la recta y la curva de distribución es el punto B. La recta se define como: $v = (s_v/s_k)(k - \bar{k}) + \bar{v}$.

El punto B, el cual tiene coordenadas (k', v'), representa los valores ya ajustados, por lo que k' corresponderá al punto de corte del estándar de desempeño. El método asume que el grado en que los expertos están de acuerdo es proporcional a la importancia relativa que los expertos dan a las dos preguntas, de ahí que se utilice una línea recta con pendiente s_v/s_k.



Escalamiento de las puntuaciones

El escalamiento (Wilson, 2005) se llevará a cabo a partir de las puntuaciones crudas de los sustentantes, y se obtendrá una métrica común para los instrumentos de evaluación, que va de 60 a 170 puntos, aproximadamente, ubicando el primer punto de corte (nivel de desempeño NII) para los instrumentos en los **100 puntos**. El escalamiento consta de dos transformaciones:

a) Transformación doble arcoseno que estabiliza la magnitud de la precisión que se tiene para cada punto de la escala.

b) Transformación lineal que ubica el primer punto de corte en 100 unidades y define el número de distintos puntos en la escala (el rango de las puntuaciones) con base en la confiabilidad del instrumento, por lo que, a mayor confiabilidad, habrá más puntos en la escala (Shun-Wen Chang, 2006).

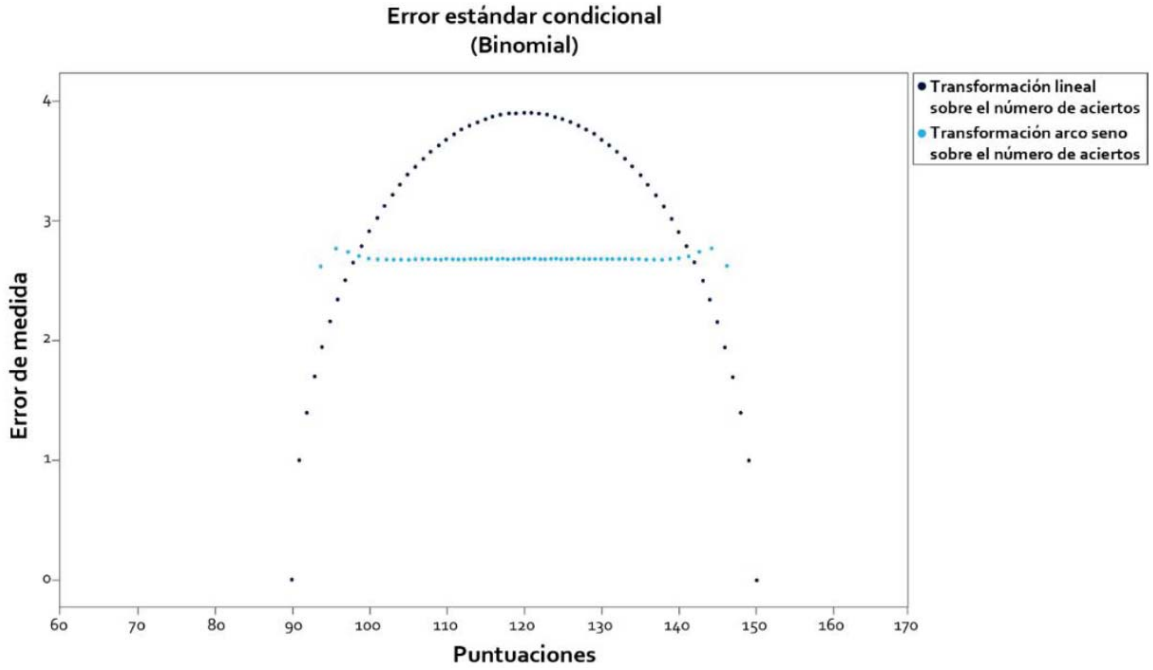
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

Finalmente, es importante destacar que para que se lleve a cabo el escalamiento, el sustentante debió alcanzar, al menos, un acierto en el instrumento de evaluación en cuestión. De no ser así, se reportará como cero y el resultado será NI.

Procedimiento para la transformación doble arcoseno

En los casos de los exámenes de opción múltiple, deberá calcularse el número de respuestas correctas que haya obtenido cada sustentante en el instrumento de evaluación. Los reactivos se calificarán como correctos o incorrectos de acuerdo con la clave de respuesta correspondiente. Si un sustentante no contesta un reactivo o si selecciona más de una alternativa de respuesta para un mismo reactivo, se calificará como incorrecto. Cuando los instrumentos de evaluación sean calificados por rúbricas, deberá utilizarse el mismo procedimiento para asignar puntuaciones a los sustentantes considerando que K sea la máxima puntuación que se pueda obtener en el instrumento de evaluación.

Cuando se aplica la transformación doble arcoseno sobre el número de aciertos obtenido en el instrumento de evaluación, el error estándar condicional de medición de las puntuaciones obtenidas se estabiliza, es decir, es muy similar, pero no igual, a lo largo de la distribución de dichas puntuaciones, con excepción de los valores extremos, a diferencia de si se aplica una transformación lineal, tal y como se observa en la siguiente gráfica (Won-Chan, Brennan y Kolen, 2000).



Para estabilizar la varianza de los errores estándar condicionales de medición a lo largo de la escala y por tanto medir con similar precisión la mayoría de los puntajes de la escala, se utilizará la función *c:*

$$c(k_i) = \frac{1}{2}\left\{\operatorname{arcsen}\sqrt{\frac{k_i}{K+1}} + \operatorname{arcsen}\sqrt{\frac{k_i+1}{K+1}}\right\} \quad (1)$$

Donde:

i se refiere a un sustentante

k_i es el número de respuestas correctas que el sustentante *i* obtuvo en el instrumento de evaluación

K es el número de reactivos del instrumento de evaluación

Procedimiento para la transformación lineal

Una vez que se aplica la transformación doble arcoseno que estabiliza la magnitud de la precisión que se tiene para cada punto de la escala, se procede a aplicar la transformación lineal que ubica el primer punto de corte en 100 unidades.

La puntuación mínima aceptable que los sustentantes deben tener para ubicarse en el nivel de desempeño NII en los instrumentos de evaluación, se situará en el valor 100. Para determinarla se empleará la siguiente ecuación:

$$P_i = A * c(k_i) + B \quad (2)$$

Donde $A = \frac{Q}{[c(K)-c(0)]}$, $B = 100 - A * c(PC1)$, *Q* es la longitud de la escala, *c(K)* es la función *c* evaluada en *K*, *c(0)* es la misma función *c* evaluada en cero y PC1 es el primer punto de corte (en número de aciertos) que se definió para establecer los niveles de desempeño y que corresponde al mínimo número de aciertos que debe tener un sustentante para ubicarlo en el nivel de desempeño NII.

El valor de Q dependerá de la confiabilidad del instrumento. Para confiabilidades iguales o mayores a 0.90, Q tomará el valor 80 y, si es menor a 0.90 tomará el valor 60 (Kolen y Brennan, 2014). Lo anterior implica que los extremos de la escala pueden tener ligeras fluctuaciones.

Por último, las puntuaciones P_i deben redondearse al entero más próximo, utilizando el criterio de que puntuaciones con cinco décimas o más, suben al siguiente entero.

Cálculo de las puntuaciones de los contenidos específicos de primer nivel en los instrumentos de evaluación

Para calcular las puntaciones del sustentante (*i*) en los contenidos específicos del primer nivel, se utilizará la puntuación ya calculada para el examen (P_i), el número de aciertos de todo el instrumento de evaluación (k_i), y el número de aciertos de cada uno de los contenidos específicos que conforman el instrumento (k_{Aji}). Las puntuaciones de los contenidos específicos (P_{Aji}) estarán expresadas en números enteros y su suma deberá ser igual a la puntuación total del instrumento (P_i).

Si el instrumento de evaluación está conformado por dos contenidos específicos, primero se calculará la puntuación del contenido específico que tenga la mayor relevancia (P_{A1i}), de acuerdo con la jerarquización de los contenidos de primer nivel (que se indicaron en el apartado de la definición de las listas de prelación), mediante la ecuación:

$$P_{A1i} = P_i * \frac{k_{A1i}}{k_i} \quad (3)$$

El resultado se redondeará al entero inmediato anterior con el criterio de que puntuaciones con cinco décimas o más suben al siguiente entero. La otra puntuación del contenido específico del primer nivel (P_{A2i}) se calculará como:

$$P_{A2i} = P_i - P_{A1i} \quad (4)$$

Para los instrumentos de evaluación con más de dos contenidos específicos, se calculará la puntuación de cada uno siguiendo el mismo procedimiento empleando la ecuación (3) para los primeros. La puntuación del último contenido específico, que tiene una menor prioridad, se calculará por sustracción como complemento de la puntuación del instrumento de evaluación, el resultado se redondeará al entero positivo más próximo. De esta manera, si el instrumento consta de *j* contenidos específicos, la puntuación del *j-ésimo* contenido específico será:

$$P_{Aji} = P_i - \sum_{k=1}^{j-1} P_{Aki} \quad (5)$$

En los casos donde el número de aciertos de un conjunto de contenidos específicos del instrumento sea cero, no se utilizará la fórmula (3) debido a que no está definido el valor de un cociente en donde el denominador tome el valor de cero. En este caso, el puntaje deberá registrarse como cero.

Procedimiento para el error estándar condicional. Método delta

Dado que el error estándar de medición se calcula a partir de la desviación estándar de las puntuaciones y su correspondiente confiabilidad, dicho error es un 'error promedio' de todo el instrumento. Por lo anterior, se debe implementar el cálculo del error estándar condicional de medición (CSEM), que permite evaluar el error estándar de medición (SEM) para puntuaciones específicas, por ejemplo, los puntos de corte.

Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta, (Muñiz, 2003), que calcula los errores estándar de medición condicionales. Para incluir la confiabilidad del instrumento de medición se usa un modelo de error binomial, para el cálculo del error estándar condicional de medición será:

$$\boldsymbol{\sigma(X) = \sqrt{\frac{1-\alpha}{1-KR21}\left[\frac{X(n-X)}{n-1}\right]}}$$

Donde:

X es una variable aleatoria asociada a los puntajes

n es el número de reactivos del instrumento

KR21 es el coeficiente de Kuder-Richardson.

α es el coeficiente de confiabilidad de Cronbach, KR-20 (Thompson, 2003):

$$\boldsymbol{\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n}\sigma_J^2}{\sigma_X^2}\right)}$$

$\sum_{j=1}^{n}\sigma_J^2$ = suma de las varianzas de los n reactivos

σ_X^2 = varianza de las puntuaciones en el instrumento

Para calcular el error estándar condicional de medición de la transformación P_i, se emplea el Método delta, el cual establece que si $P_i = g(X)$, entonces un valor aproximado de la varianza de $g(X)$ está dado por:

$$\boldsymbol{\sigma^2(P_i) \doteq \left(\frac{dg(X)}{dX}\right)^2 \sigma^2(X)}$$

De ahí que:

$$\boldsymbol{\sigma(P_i) \doteq \frac{dg(x)}{dx}\,\sigma(x)}$$

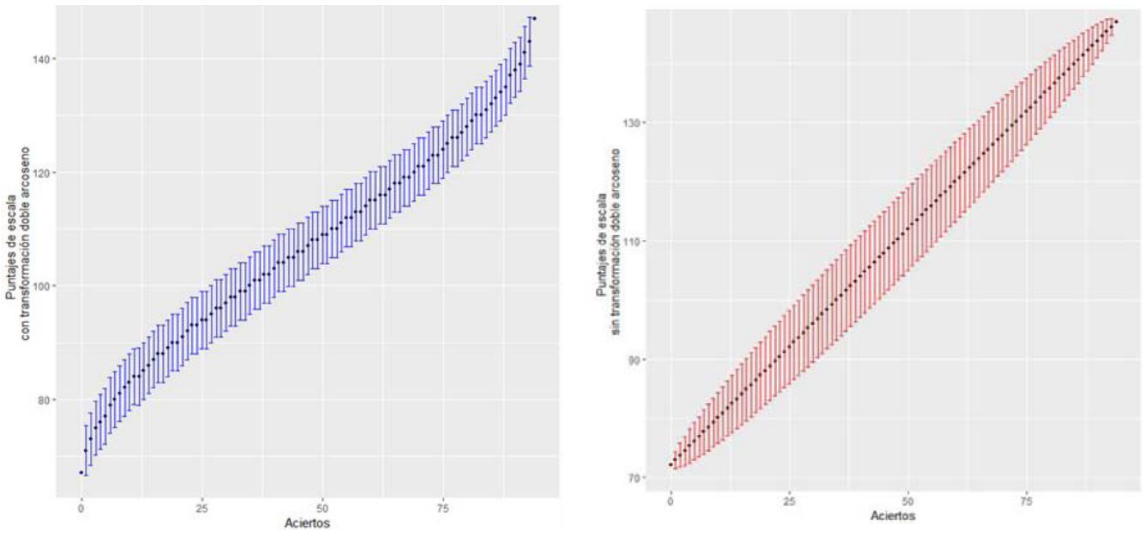
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

$$\boldsymbol{\sigma(P_i) \doteq \frac{A}{2}\left[\frac{1}{2(k+1)\left(\sqrt{\frac{x}{k+1}}\right)\left(\sqrt{1-\frac{x}{k+1}}\right)} + \frac{1}{2(k+1)\left(\sqrt{\frac{x+1}{k+1}}\right)\left(\sqrt{1-\frac{x+1}{k+1}}\right)}\right]\sigma(x)}$$

Donde $\sigma(x)$ es el error estándar de medida de las puntuaciones crudas y $\sigma(P_i)$ el error estándar condicional de medición, de la transformación P_i, que ya incorpora la confiabilidad.

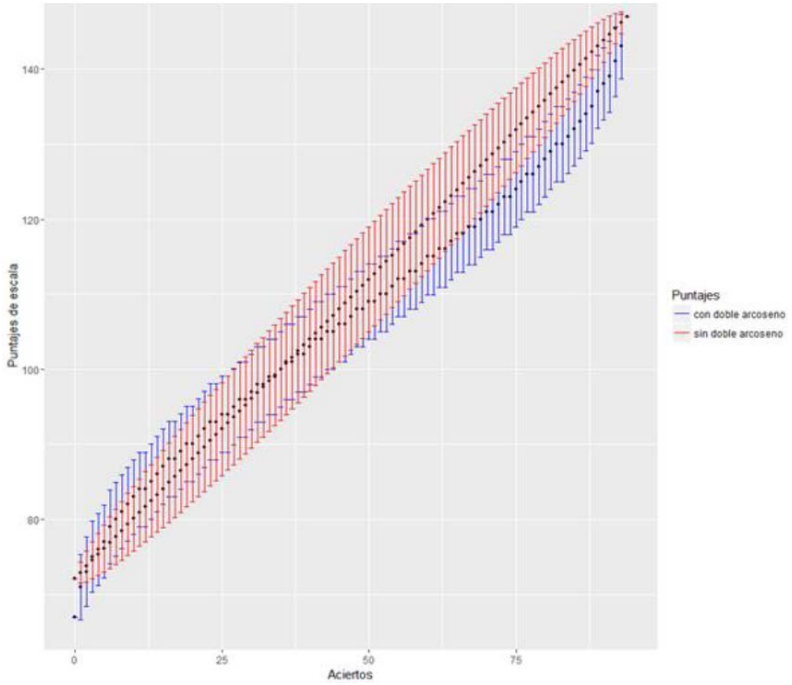
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 2012; American College Testing, 2013; 2014a; 2014b).

En las siguientes gráficas se muestran los intervalos de confianza (al 95% de confianza) de los puntajes de la escala cuando se aplica la transformación doble arcoseno (gráfica del lado izquierdo) y cuando no se aplica (gráfica del lado derecho).



Se observa que al aplicar la transformación doble arcoseno se mide con similar precisión la mayoría de los puntajes de la escala, a diferencia de cuando no se aplica dicha transformación, además de que en el punto de corte para alcanzar el nivel de desempeño NII (100 puntos) el error es menor cuando se aplica la transformación.

Esto es más claro si se observan ambas gráficas en el mismo cuadrante, como en la siguiente imagen.



El dato obtenido del error estándar condicional deberá reportarse en la misma escala en que se comunican las calificaciones de los sustentantes e incorporarse en el informe o manual técnico del instrumento (estándar 2.13 de los Estándares para las Pruebas Educativas y Psicológicas de la *American Educational Research Association* et. al., 2014). Asimismo, esto permite atender al estándar 2.14 de los Estándares para las Pruebas Educativas y Psicológicas de la *American Educational Research Association* et. al., 2014, el cual establece que cuando se especifican puntos de corte para selección o clasificación, los errores estándar deben ser reportados en la vecindad de cada punto de corte en dicho informe o manual técnico.

Proceso para la equiparación de instrumentos de evaluación

El procedimiento que permite hacer equivalentes los resultados obtenidos en diferentes formas o versiones de un mismo instrumento es una equiparación. La que aquí se plantea considera dos estrategias: a) si el número de sustentantes es de al menos 100 en ambas formas, se utilizará el método de equiparación lineal de Levine para puntajes observados; o bien, b) si el número de sustentantes es menor de 100 en alguna de las formas, se utilizará el método de equiparación de identidad (*identity equating*). A continuación, se detallan los procedimientos.

Método de equiparación lineal de Levine

La equiparación de las formas de un instrumento deberá realizarse utilizando el método de equiparación lineal de Levine (Kolen y Brennan, 2014) para puntajes observados bajo un diseño de grupos no equivalentes con reactivos comunes. Dicho diseño es uno de los más utilizados en la práctica. En cada muestra de sujetos se administra solamente una forma de la prueba, con la peculiaridad de que en ambas muestras se administra un conjunto de reactivos en común llamado ancla, que permite establecer la equivalencia entre las formas a equiparar.

Cualquiera de los métodos de equiparación de puntajes que se construya involucra dos poblaciones diferentes. Sin embargo, una función de equiparación de puntajes se define sobre una población única. Por lo tanto, las poblaciones 1 y 2 que corresponden a las poblaciones donde se aplicó la forma nueva y antigua, deben ser combinadas para obtener una población única a fin de definir una relación de equiparación.

Esta única población se conoce como población sintética, en la cual se le asignan pesos w_1 y w_2 a las poblaciones 1 y 2, respectivamente, esto es, $\mathbf{w_1 + w_2 = 1}$ y $\mathbf{w_1, w_2 \geq 0}$. Para este proceso se utilizará

$$\boldsymbol{w_1 = \frac{N_1}{N_1 + N_2}}$$

y

$$\boldsymbol{w_2 = \frac{N_2}{N_1 + N_2}}$$

Donde N_1 corresponde al tamaño de la población 1 y N_2 corresponde al tamaño de la población 2.

Los puntajes de la forma nueva, aplicada a la población 1, serán denotados por X; los puntajes de la forma antigua, aplicada a la población 2, serán denotados por Y.

Los puntajes comunes están identificados por V y se dice que los reactivos comunes corresponden a un anclaje interno cuando V se utiliza para calcular los puntajes totales de ambas poblaciones.

Usando el concepto de población sintética, la relación lineal de equiparación de puntajes para el diseño de grupos no equivalentes con reactivos comunes se escribe de la siguiente forma:

$$\boldsymbol{l_{Y_s}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y)}$$

Donde s denota la población sintética y

$$\boldsymbol{\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)]}$$

$$\boldsymbol{\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)]}$$

$$\boldsymbol{\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_1^2[\mu_1(V) - \mu_2(V)]^2}$$

$$\boldsymbol{\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_2^2[\mu_1(V) - \mu_2(V)]^2}$$

Donde los subíndices 1 y 2 se refieren a las poblaciones 1 y 2 respectivamente.

$$\boldsymbol{\gamma_1 = \frac{\sigma_1(X, V)}{\sigma_1^2(V)}}$$

y

$$\gamma_2 = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}$$

Específicamente, para el método de Levine para puntajes observados bajo un diseño de grupos no equivalentes con reactivos comunes, las $\gamma's$ se expresan de la siguiente manera:

$$\gamma_1 = \frac{\sigma_1^2(X)}{\sigma_1(X,V)}$$

$$\gamma_2 = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)}$$

Para aplicar este método basta con reemplazar estos coeficientes en las ecuaciones lineales antes descritas. Por su parte, Kolen y Brennan proveen justificaciones para usar esta aproximación.

Es importante señalar que para los puntajes que se les aplique la equiparación $x_e = b_1 x + b_0$, con b_1 como pendiente y b_0 como ordenada al origen, el procedimiento es análogo al descrito en la sección "Procedimiento para el error estándar condicional. Método delta", y el error estándar condicional de medición para la transformación $P_{i_e} = A * c(x_e) + B$, que ya incorpora la confiabilidad, está dado por:

$$\sigma(P_{i_e}) \doteq \frac{A}{2}\left[\frac{b1}{2(k+1)\left(\sqrt{\frac{x_e}{k+1}}\right)\left(\sqrt{1-\frac{x_e}{k+1}}\right)} + \frac{b1}{2(k+1)\left(\sqrt{\frac{x_e+1}{k+1}}\right)\left(\sqrt{1-\frac{x_e+1}{k+1}}\right)}\right]\sigma(x_e)$$

Donde x_e son las puntuaciones equiparadas, las cuales son una transformación de las puntuaciones crudas, por lo que el error estándar de medida de dicha transformación se define como:

$$\sigma(x_e) = b_1 * \sigma(x)$$

Método de equiparación de identidad (*identity equating*)

La equiparación de identidad es la más simple, toda vez que no hace ningún ajuste a la puntuación "*x*" en la escala de la forma *X* al momento de convertirla en la puntuación equiparada "y" en la escala de la forma *Y*.

Es decir, dichas puntuaciones son consideradas equiparadas cuando tienen el mismo valor, por lo que las coordenadas de la línea de equiparación de identidad están definidas simplemente como *x=y* (Holland y Strawderman, 2011).

Referencias

American College Testing. (2013). *ACT Plan Assessments Technical Manual*, Iowa City, IA: Author.

American College Testing. (2014a) *ACT Assessments Technical Manual*, Iowa City, IA: Author.

American College Testing. (2014b) *ACT QualityCore Assessments Technical Manual*, Iowa City, IA: Author.

AERA, APA y NCM. (2014). S*tandards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Angoff, W. (1971). *Scales, norms, and equivalent scores*. En R. Thorndike (Ed.), Educational Measurement (2° ed.). Washington, DC: American Council on Education.

Beuk, C. H. (1984). A Method for Reaching a Compromise between Absolute and Relative Standards in Examinations. *Journal of Educational Measurement*, 21 (2) pp. 147-152.

Brennan, R. L. (2012). Scaling PARCC Assessments: Some considerations and a synthetic data example.

Cook, D. y Beckman, T. (2006). *Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application*. *The American Journal of Medicine 119*, 166.e7-166.e16

Downing, S. (2004). Reliability: On the reproducibility of assessment data. Med Educ 38(9) pp.1006-1012.

Holland, P. y Strawderman, W. (2011). How to average equating functions, if you must. En Von Davier, A. (Ed.). *Statistical models for test equating, scaling, and linking* (pp. 89–107). Nueva York, NY: Springer.

Jonsson, A. y Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2 pp.130-144.

Kendall, M. y Stuart, A. (1977). *The advanced theory of statistics, Vol. 1: Distribution theory*. 4ª Ed. New York, NY: MacMillan.

Kolen, M. y Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices*. 3ª Ed. New York, NY: Springer-Verlag.

Masters, G. (1982). *A Rasch model for Partial Credit Scoring*. Psychometrika. 47(2) pp 149-174.

Muñiz, J. (2003). *Teoría clásica de los test.* Madrid, España: Ediciones pirámide.

Muraki, E. (1999*). Stepwise Analysis of Differential Item Functioning Based on Multiple-Group Partial Credit Model.* Journal of Educational Measurement.

OECD (2002). *PISA 2000 Technical Report.* Paris, Francia: OECD Publishing.

OECD (2005). *PISA 2003 Technical Report*. Paris, Francia: OECD Publishing.

OECD (2009). *PISA 2006 Technical Report*. Paris, Francia: OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. Paris, Francia: OECD Publishing.

Rezaei, A. y Lovorn, M. (2010). *Reliability and validity of rubrics for assessment through writing. Assessing Writing* 15 (1) 18-39.

Chang, S. (2006). *Methods in Scaling the Basic Competence Test, Educational and Psychological Measurement,* 907-927

Stellmack, M., Konheim-Kalkstein, Y., Manor, J., Massey, A. y Schmitz, J. (2009). An assessment of reliability and validity of a rubric for APA-style introductions, *Teaching of Psychology,* 36 pp. 102-107.

Stemler, E. y Tsai, J. (2008). 3 Best Practices in Interrater Reliability Three Common Approaches*.* En *Best practices in quantitative methods.* pp. 29-49. California, E.U.A.: SAGE Publications, Inc.

Thompson, B. (ed.). (2003). *Score reliability. Contemporary thinking on reliability issues*. California, E.U.A.: SAGE Publications, Inc.

Wilson, M. (2005). *Constructing measures. An ítem response modeling approach*. Mahwah, Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

Won-Chan, L., Brennan, R. y Kolen, M. (2000). Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study. *Journal of Educational Measurement*, 37 (1) pp. 1-20.

Wu, M. y Adams, R. (2007). *Applying the Rasch Model to Psycho-social measurement. A practical approach.* Melbourne, Australia: Educational measurement solutions.

TRANSITORIOS

Primero. Los presentes Criterios entrarán en vigor al día siguiente de su publicación en el Diario Oficial de la Federación.

Segundo. Los presentes Criterios, de conformidad con los artículos 40 y 48 de la Ley del Instituto Nacional para la Evaluación de la Educación, deberán hacerse del conocimiento público a través de la página de Internet del Instituto www.inee.edu.mx

Tercero. Se instruye a la Dirección General de Asuntos Jurídicos para que realice las gestiones necesarias a efecto de que los presentes Criterios se publiquen en el Diario Oficial de la Federación.

Ciudad de México, a trece de marzo de dos mil diecinueve.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Cuarta Sesión Extraordinaria de dos mil diecinueve, celebrada el trece de marzo de dos mil diecinueve.- Acuerdo número SEJG/04-19/03,R. La Consejera Presidenta, **Teresa Bracho González**.- Los Consejeros: **Bernardo Hugo Naranjo Piñera, Sylvia Irene Schmelkes del Valle** y **Patricia Gabriela Vázquez del Mercado Herrera.**

El Director General de Asuntos Jurídicos, **Agustín E. Carrillo Suárez**.- Rúbrica.

(R.- 480010)

CRITERIOS técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la definición de las listas de prelación de los Concursos de oposición para las promociones a cargos con funciones de dirección y supervisión en Educación Básica y Educación Media Superior y a las funciones de asesoría técnica pedagógica en Educación Básica para el ciclo escolar 2019-2020.

Al margen un logotipo, que dice: Instituto Nacional para la Evaluación de la Educación.- México.

CRITERIOS TÉCNICOS Y DE PROCEDIMIENTO PARA EL ANÁLISIS DE LOS INSTRUMENTOS DE EVALUACIÓN, EL PROCESO DE CALIFICACIÓN Y LA DEFINICIÓN DE LAS LISTAS DE PRELACIÓN DE LOS CONCURSOS DE OPOSICIÓN PARA LAS PROMOCIONES A CARGOS CON FUNCIONES DE DIRECCIÓN Y SUPERVISIÓN EN EDUCACIÓN BÁSICA Y EDUCACIÓN MEDIA SUPERIOR Y A LAS FUNCIONES DE ASESORÍA TÉCNICA PEDAGÓGICA EN EDUCACIÓN BÁSICA PARA EL CICLO ESCOLAR 2019-2020

Con fundamento en los Lineamientos para llevar a cabo la evaluación para la promoción de docentes a cargos con funciones de Dirección, Supervisión y Asesoría Técnica Pedagógica en Educación Básica y para la promoción de docentes a cargos con funciones de Dirección y Supervisión en Educación Media Superior para el ciclo escolar 2019-2020. LINEE-02-2019; la Junta de Gobierno aprueba emitir los siguientes Criterios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la definición de las listas de prelación de los Concursos de oposición para las promociones a cargos con funciones de dirección y supervisión en Educación Básica y Educación Media Superior y a las funciones de asesoría técnica pedagógica en Educación Básica para el ciclo escolar 2019-2020.

Estos criterios técnicos se deben adoptar en los siguientes concursos de oposición.

Educación Básica:

- Concursos para la promoción a cargos con funciones de dirección: Director, Subdirector y Coordinador de actividades.
- Concursos para la promoción a cargos con funciones de supervisión: Jefe de sector, Jefe de enseñanza y Supervisor.
- Concursos para la promoción a funciones de Asesoría Técnica Pedagógica (ATP): ATP en Lenguaje oral y escrito, ATP en Pensamiento matemático, ATP en Educación especial y ATP en Educación física.

Educación Media Superior:

- Concursos para la promoción a cargos con funciones de dirección: Director, Subdirector académico, Subdirector administrativo, Jefe de departamento académico y Jefe de departamento administrativo.
- Concursos para la promoción a cargos con funciones de supervisión.

Los presentes Criterios técnicos y de procedimiento consideran el uso de los datos recabados una vez que se ha llevado a cabo la aplicación de los instrumentos que forman parte de la evaluación y tienen como finalidad establecer los referentes necesarios para garantizar la validez, confiabilidad y equidad de los resultados de los procesos de evaluación implicados en estos concursos. Su contenido se organiza en cinco apartados: 1) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 2) Procedimiento para el establecimiento de puntos de corte y estándares de desempeño; 3) Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos del proceso de evaluación; 4) Resultado del proceso de evaluación: conformación de los grupos de desempeño y 5) Integración de las listas de prelación.

En la parte final, se integra un Anexo técnico con información detallada de los aspectos técnicos que se consideran en los distintos apartados del documento.

Definición de términos

Para los efectos del presente documento, se emplean las siguientes definiciones:

I. **Alto impacto:** Se entiende que una evaluación es de alto impacto cuando sus resultados tienen consecuencias importantes para las personas o las instituciones; por ejemplo, los procesos de admisión o certificación.

II. **Calificación:** Proceso de asignación de una puntuación o nivel de desempeño logrado a partir de los resultados de una medición.

III. **Confiabilidad:** Cualidad de las mediciones obtenidas con un instrumento, que se caracterizan por ser consistentes y estables cuando este se aplica en distintas ocasiones.

IV. **Constructo:** Elaboración teórica formulada para explicar un proceso social, psicológico o educativo.

V. **Correlación punto biserial:** Medida de consistencia que se utiliza en el análisis de reactivos, indica si hay una correlación entre el resultado de un reactivo con el resultado global del examen.

VI. **Criterio de evaluación:** Indicador de un valor aceptable sobre el cual se puede establecer o fundamentar un juicio de valor sobre el desempeño de una persona.

VII. **Desempeño:** Resultado obtenido por el sustentante en un proceso de evaluación o en un instrumento de evaluación educativa.

VIII. **Dificultad de un reactivo:** Indica la proporción de personas que responden correctamente el reactivo de un examen.

IX. **Distractores:** Opciones de respuesta incorrectas del reactivo de opción múltiple, que probablemente serán elegidas por los sujetos con menor dominio en lo que se evalúa.

X. **Dominio:** Conjunto de conocimientos, habilidades, destrezas, actitudes u otros atributos que tienen las siguientes propiedades: límites, extensión y definición. También se puede aplicar a contenidos, procedimientos u objetos.

XI. **Educación básica:** Tipo de educación que comprende los niveles de preescolar, primaria y secundaria en todas sus modalidades, incluyendo la educación indígena, la especial y la que se imparte en los centros de educación básica para adultos.

XII. **Educación media superior:** Tipo de educación que comprende el nivel de bachillerato, los demás niveles equivalentes a éste, así como la educación profesional que no requiere bachillerato o sus equivalentes.

XIII. **Equiparación:** Método estadístico que se utiliza para ajustar las puntuaciones de las formas o versiones de un mismo instrumento, de manera tal que al sustentante le sea indistinto, en términos de la puntuación que se le asigne, responder una forma u otra.

XIV. **Error estándar de medida:** Es la estimación de mediciones repetidas de un atributo evaluado sobre una misma persona a partir de un mismo instrumento y que tienden a distribuirse alrededor de un puntaje verdadero, el cual siempre es desconocido porque ninguna medida lo representa a la perfección.

XV. **Escala:** Conjunto de números, puntuaciones o medidas que pueden ser asignados a objetos o sucesos con propiedades específicas a partir de reglas definidas.

XVI. **Escalamiento:** Proceso a través del cual se construye una escala que facilita la interpretación de los resultados que se obtienen en uno o varios instrumentos de evaluación, colocando las puntuaciones de los distintos instrumentos o formas a una escala común.

XVII. **Especificaciones de tareas evaluativas o de reactivos:** Descripción detallada de las tareas específicas susceptibles de medición, que deben realizar las personas que contestan el instrumento de evaluación. Deben estar alineadas al constructo definido en el marco conceptual.

XVIII. **Estándar:** Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación. Constituyen el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XIX. **Estructura:** Está compuesta por los contenidos o aspectos disciplinares que mide un instrumento, así como el número y porcentaje relativo de reactivos o tareas evaluativas con que se integra el instrumento y su distribución.

XX. **Evaluación:** Proceso sistemático mediante el cual se recopila y analiza información, cuantitativa o cualitativa, sobre un objeto, sujeto o evento, con el fin de emitir juicios de valor al comparar los resultados con un referente previamente establecido. La información resultante puede ser empleada como insumo para orientar la toma de decisiones.

XXI. **Examen:** Instrumento de evaluación que se emplea para identificar el nivel de dominio de los sustentantes sobre un constructo específico.

XXII. **Formas de un instrumento**: Dos o más versiones de un instrumento que se consideran equivalentes, pues se construyen con los mismos contenidos y especificaciones estadísticas.

XXIII. **Instrumento de evaluación:** Herramienta de recolección de datos que suele tener distintos formatos, atendiendo a la naturaleza de la evaluación, por ejemplo, instrumentos de selección de respuesta, instrumentos de respuesta construida, cuestionarios, observaciones, portafolios, entre otros.

XXIV. **Instrumento de evaluación referido a un criterio:** Instrumento que permite comparar el desempeño de las personas evaluadas, con un estándar preestablecido.

XXV. **Jueceo:** Método en el cual se utiliza la opinión de expertos (denominados jueces) para valorar y calificar distintos aspectos, tales como las respuestas y ejecuciones de las personas que participan en una evaluación o la calidad de los reactivos, las tareas evaluativas y estándares de un instrumento.

XXVI. **Lista de prelación:** Orden descendente en que se enlistan los sustentantes con base en los resultados obtenidos en el proceso de evaluación**.**

XXVII. **Medición:** Proceso de asignación de valores numéricos a atributos de las personas, características de objetos o eventos de acuerdo con reglas específicas que permitan que sus propiedades puedan ser representadas cuantitativamente.

XXVIII. **Muestra:** Subconjunto de la población de interés que refleja las variables medidas en una distribución semejante a la de la población.

XXIX. **Multi-reactivo:** Conjunto de reactivos de opción múltiple que están vinculados a un planteamiento general, por lo que este último es indispensable para poder resolverlos.

XXX. **Nivel de desempeño:** Criterio conceptual que delimita el marco interpretativo de las puntuaciones obtenidas en una prueba y que refiere a lo que el sustentante es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.

XXXI. **Objeto de medida:** Conjunto de características o atributos que se miden en el instrumento de evaluación.

XXXII. **Parámetro estadístico:** Número que resume un conjunto de datos que se derivan del análisis de una cualidad o característica del objeto de estudio.

XXXIII. **Perfil:** Conjunto de características, requisitos, cualidades o aptitudes que deberá tener el sustentante a desempeñar un puesto o función descrito específicamente.

XXXIV. **Porcentaje de acuerdos inter-jueces:** Medida del grado en que dos jueces coinciden en la puntuación asignada a un producto o proyecto definido para evaluar, a través de una rúbrica, el desempeño de una persona.

XXXV. **Porcentaje de acuerdos intra-jueces:** Medida del grado en que el mismo juez, a través de dos o más mediciones repetidas a los mismos productos o proyectos evaluados con la misma rúbrica, coincide en las puntuaciones que les asignó.

XXXVI. **Protocolo:** Conjunto de reglas o normas que establecen cómo se deben realizar ciertas actividades o pasos, además de incorporar de manera detallada los procedimientos y los estándares que se deben cumplir. Para su realización es fundamental considerar las características especiales de cada evaluación.

XXXVII. **Punto de corte:** En instrumentos de evaluación con referencia a un estándar de desempeño, es la puntuación mínima o el criterio a alcanzar o a superar para considerar que el nivel de desempeño de una persona cumple con lo esperado y distinguirlo de otro que no.

XXXVIII. **Puntuación:** Valor numérico obtenido durante el proceso de medición.

XXXIX. **Reactivo:** Unidad básica de medida de un instrumento de evaluación que consiste en una pregunta o instrucción que requiere una respuesta del sujeto.

XL. **Rúbrica:** Herramienta que integra los criterios de ejecución a partir de los cuales se califica una tarea evaluativa.

XLI. **Sesgo:** Error en la medición de un atributo debido a una variable no controlada, como las diferencias culturales o lingüísticas de las personas evaluadas.

XLII. **Tarea evaluativa:** Unidad básica de medida de un instrumento de evaluación que consiste en la respuesta que construye una persona o en la ejecución de una actividad, que es susceptible de ser observada y graduada en su nivel de cumplimiento.

XLIII. **Teoría Clásica de los Tests (TCT):** Teoría psicométrica que parte del supuesto de que el puntaje observado de una persona que responde un instrumento es la suma de su puntaje verdadero y un error aleatorio independiente del puntaje.

XLIV. **Validez:** Juicio valorativo integrador sobre el grado en que los fundamentos teóricos y las evidencias empíricas apoyan la interpretación de las puntuaciones de los instrumentos de evaluación.

1. Criterios técnicos para el análisis e integración de los instrumentos de evaluación

Uno de los aspectos fundamentales que debe llevarse a cabo antes de emitir cualquier resultado de un proceso de evaluación es el análisis psicométrico de los instrumentos que integran la evaluación, con el objetivo de verificar que cuentan con la calidad técnica necesaria para proporcionar resultados confiables, acordes con el objetivo de la evaluación.

Las técnicas empleadas para el análisis de un instrumento dependen de su naturaleza, de los objetivos específicos para los cuales fue diseñado, así como del tamaño de la población evaluada. Sin embargo, en todos los casos, debe aportarse información sobre la dificultad y discriminación de sus reactivos o tareas evaluativas, así como la precisión del instrumento, los indicadores de consistencia interna o estabilidad del instrumento, los cuales, además de los elementos asociados a la conceptualización del objeto de medida, forman parte de las evidencias que servirán para valorar la validez de la interpretación de sus resultados. Estos elementos, deberán reportarse en el informe o manual técnico del instrumento.

Con base en los resultados de estos procesos de análisis deben identificarse las tareas evaluativas o los reactivos que cumplen con los criterios psicométricos especificados en este documento para integrar el instrumento, para calificar el desempeño de las personas evaluadas, con la mayor precisión posible.

Para llevar a cabo el análisis de los instrumentos de medición utilizados en el proceso de evaluación, es necesario que los distintos grupos de sustentantes de las entidades federativas queden equitativamente representados, dado que la cantidad de sustentantes por tipo de evaluación en cada entidad federativa es notoriamente diferente. Para ello, se definirá una muestra de sustentantes por cada instrumento de evaluación que servirá para analizar el comportamiento estadístico de los instrumentos y orientar los procedimientos descritos más adelante, y que son previos para la calificación. Para conformar dicha muestra, cada entidad federativa contribuirá con 500 sustentantes como máximo, y deberán ser elegidos aleatoriamente. Si hay menos de 500 sustentantes, todos se incluirán en la muestra (OECD; 2002, 2005, 2009, 2014). Si no se realizara este procedimiento, las decisiones sobre los instrumentos de evaluación, la identificación de los puntos de corte y los estándares de desempeño, se verían fuertemente influenciados, indebidamente, por el desempeño mostrado por aquellas entidades que se caracterizan por tener un mayor número de sustentantes.

Sobre la conformación de los instrumentos de evaluación

Con la finalidad de obtener puntuaciones de los aspirantes con el nivel de precisión requerido para los propósitos de los Concursos, los instrumentos de evaluación deberán tener las siguientes características:

Exámenes con reactivos de opción múltiple:

Los instrumentos de evaluación deberán tener, al menos, 80 reactivos efectivos para calificación y deberán estar organizados jerárquicamente en tres niveles de desagregación: áreas, subáreas y temas, en donde:

- Las áreas deberán contar con al menos dos subáreas y, cada una de ellas, deberá tener al menos 20 reactivos efectivos para calificar.
- Las subáreas deberán considerar al menos dos temas, y cada uno de ellos deberá tener, al menos, 10 reactivos efectivos para calificar.
- Los temas deberán contemplar al menos dos contenidos específicos, los cuales estarán definidos en términos de especificaciones de reactivos. Cada especificación deberá ser evaluada al menos por un reactivo.

Exámenes de respuesta construida:

- Deberán estar organizados en, al menos, dos niveles de desagregación (áreas y subáreas; si fuera el caso, temas); el primero deberá contar, al menos, con dos conjuntos de contenidos específicos a evaluar.
- A partir del segundo nivel (o tercer nivel, si fuera el caso) de desagregación, se deberá contar con las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional.
- En las rúbricas o guías de calificación los distintos niveles o categorías de ejecución que se consignen, deberán ser claramente distinguibles entre sí y con un diseño ordinal ascendente (de menor a mayor valor).

Criterios y parámetros estadísticos

Los instrumentos empleados para los Concursos de oposición para la Promoción en Educación Básica y en Educación Media Superior para el ciclo escolar 2019-2020, deberán atender los siguientes criterios (Cook y Beckman 2006; Downing, 2004; Stemler y Tsai, 2008) con, al menos, los valores de los parámetros estadísticos indicados a continuación:

En el caso de los instrumentos de evaluación basados en reactivos de opción múltiple:

- La respuesta correcta deberá tener una dificultad clásica de 10% a 90% y una correlación punto biserial corregida igual o mayor que 0.15.
- Los distractores deberán tener correlaciones punto biserial negativas.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.80.

Para los instrumentos con menos de 100 sustentantes, la selección de los reactivos con los cuales se va a calificar, se debe llevar a cabo con base en el siguiente procedimiento: cada reactivo tiene que ser revisado por, al menos, tres jueces: dos expertos en contenido y un revisor técnico, considerando los siguientes aspectos: *calidad del contenido del reactivo, adecuada construcción técnica, correcta redacción y atractiva presentación de lo que se evalúa*.

En todos los casos en los que sea factible estimar los parámetros estadísticos de los reactivos, esta información debe proporcionarse a los jueces con el objetivo de que les permita fundamentar sus decisiones y ejercer su mejor juicio profesional.

En el caso de los instrumentos basados en tareas evaluativas o en reactivos de respuesta construida, y que serán calificados con rúbrica:

- La correlación corregida entre cada aspecto evaluado con la puntuación global deberá ser igual o mayor que 0.20.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.80.

Considerando las decisiones de los jueces que calificaron los instrumentos de respuesta construida a través de rúbrica, se debe atender lo siguiente:

- El porcentaje de acuerdos inter-jueces deberá ser igual o mayor que 60%.
- El porcentaje de acuerdos intra-jueces deberá ser igual o mayor que 60% considerando, al menos, cinco medidas repetidas seleccionadas al azar, es decir, para cada juez se deben seleccionar al azar cinco sustentantes, a quienes el juez debe calificar en dos ocasiones a través de la estrategia "doble ciego", a fin de que, en su segunda calificación, el juez no identifique el producto del sustentante. Estas mediciones deberán aportarse antes de emitir la calificación definitiva de los sustentantes, a fin de salvaguardar la confiabilidad de la decisión.

Si se diera el caso de que en algún instrumento no se cumpliera con los criterios y parámetros estadísticos antes indicados, la Junta de Gobierno del INEE determinará lo que procede, buscando salvaguardar el constructo del instrumento que fue aprobado por el Consejo Técnico y atendiendo al marco jurídico aplicable.

2. Procedimiento para el establecimiento de puntos de corte y estándares de desempeño

Un paso crucial en el desarrollo y uso de los instrumentos de evaluación de naturaleza criterial, como es el caso de los que se utilizan para evaluar a los sustentantes de los Concursos de oposición para la Promoción en Educación Básica y Educación Media Superior, es el establecimiento de los puntos de corte que dividen el rango de calificaciones para diferenciar entre niveles de desempeño.

En los instrumentos de evaluación de tipo criterial, la calificación obtenida por cada sustentante se contrasta con un estándar de desempeño establecido por un grupo de expertos que describe el nivel de competencia requerido para algún propósito determinado, es decir, los conocimientos y habilidades que, para cada instrumento de evaluación, se consideran indispensables para un desempeño adecuado en la función. En este sentido el estándar de desempeño delimita el marco interpretativo de las puntuaciones obtenidas en un instrumento por los sustentantes.

El procedimiento para el establecimiento de puntos de corte y estándares de desempeño incluye tres fases, las cuales se describen a continuación:

Primera fase

Con el fin de contar con un marco de referencia común para los distintos instrumentos de evaluación, se deberán establecer descriptores genéricos de los niveles de desempeño que se utilizarán y **cuya única función** es orientar a los comités académicos en el trabajo del desarrollo de los descriptores específicos de cada instrumento, tales que les permita a los docentes tener claros elementos de retroalimentación para conocer sus fortalezas y áreas de oportunidad identificadas a partir de los resultados de cada instrumento sustentado.

Para todos los instrumentos se utilizarán tres niveles de desempeño posibles: Nivel I (NI), Nivel II (NII) y Nivel III (NIII). Los descriptores genéricos para cada uno de ellos se indican en la Tabla 1.

Tabla 1. Descriptores genéricos de los niveles de desempeño

Nivel de desempeño	Descriptor
Nivel I (NI)	Dominio insuficiente de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño en la función.
Nivel II (NII)	Dominio suficiente y organizado de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño en la función.
Nivel III (NIII)	Dominio suficiente y organizado de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño en la función, con *amplia capacidad para utilizarlos en diversas situaciones propias de la función.*

Segunda fase

En esta fase se establecerán los puntos de corte y deberán participar los comités académicos específicos para el instrumento de evaluación que se esté trabajando. Dichos comités se deberán conformar, en su conjunto, con especialistas que han participado en el diseño de los instrumentos y cuya pluralidad sea representativa de la diversidad cultural en que se desenvuelve la acción educativa del país. En todos los casos, sus miembros deberán ser capacitados específicamente para ejercer su mejor juicio profesional a fin de identificar cuál es la puntuación requerida para que el sustentante alcance un determinado nivel o estándar de desempeño.

Los insumos que tendrán como referentes para el desarrollo de esta actividad serán, la documentación que describe la estructura de los instrumentos, las especificaciones, los ejemplos de tareas evaluativas o de reactivos incluidos en las mismas y las rúbricas utilizadas para la calificación. En todos los casos, los puntos de corte se referirán a la ejecución típica o esperable de un sustentante hipotético, con un desempeño mínimamente aceptable, para cada uno de los niveles. Para ello, se deberá determinar, para cada tarea evaluativa o reactivo considerado en el instrumento, cuál es la probabilidad de que dicho sustentante hipotético lo responda correctamente y, con base en la suma de estas probabilidades, establecer la calificación mínima requerida o punto de corte, para cada nivel de desempeño (Angoff, 1971).

Una vez establecidos los puntos de corte que dividen el rango de calificaciones para diferenciar los niveles de desempeño en cada instrumento, se deberán describir los conocimientos y las habilidades específicos que están implicados en cada nivel de desempeño, es decir, lo que dicho sustentante conoce y es capaz de hacer.

Tercera fase

En la tercera fase se llevará a cabo un ejercicio de retroalimentación a los miembros de los comités académicos con el fin de contrastar sus expectativas sobre el desempeño de la población evaluada, con la distribución de sustentantes que se obtiene en cada nivel de desempeño al utilizar los puntos de corte definidos en la segunda fase, a fin de determinar si es necesario realizar algún ajuste en la decisión tomada con anterioridad y, de ser el caso, llevar a cabo el ajuste correspondiente.

Los jueces deberán estimar la tasa de sustentantes que se esperaría en cada nivel de desempeño y comparar esta expectativa con los datos reales de los sustentantes, una vez aplicados los instrumentos. Si las expectativas y los resultados difieren a juicio de los expertos, deberá definirse un punto de concordancia para la determinación definitiva del punto de corte asociado a cada nivel de desempeño en cada uno de los instrumentos, siguiendo el método propuesto por Beuk (1984).

Esta tercera fase se llevará a cabo solamente para aquellos instrumentos de evaluación en los que el tamaño de la población evaluada sea igual o mayor a 100 sustentantes. Si la población es menor a 100 sustentantes, los puntos de corte serán definidos de acuerdo con lo descrito en la segunda fase.

Si se diera el caso de que algún instrumento no cumpliera con el criterio de confiabilidad indicado en el apartado previo, la Junta de Gobierno del INEE determinará el procedimiento a seguir para el establecimiento de los puntos de corte correspondientes, atendiendo al marco jurídico aplicable.

3. Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos del proceso de evaluación

Todos los sustentantes que participen en los Concursos de oposición para la Promoción en Educación Básica y en Educación Media Superior para el ciclo escolar 2019-2020 deberán sustentar dos exámenes. Cada sustentante recibirá los resultados de cada uno de los exámenes que haya presentado, así como el resultado integrado de todo el proceso de evaluación.

Una vez que se han establecido los puntos de corte en cada examen, el resultado del sustentante será ubicado en uno de los tres niveles de desempeño en función de la puntuación alcanzada. Esto implica que su resultado será comparado contra el estándar previamente establecido, con independencia de los resultados obtenidos por el conjunto de sustentantes que presentaron el examen.

3.1 Proceso para la equiparación de instrumentos de evaluación

Cuando el proceso de evaluación implica la aplicación de un instrumento en diversas ocasiones en un determinado periodo, en especial si sus resultados tienen un alto impacto, es indispensable el desarrollo y uso de formas o versiones del instrumento que sean equivalentes a fin de garantizar que, independientemente del momento en que un aspirante participe en el proceso de evaluación, no tenga ventajas o desventajas de la forma o versión que responda. Por esta razón, es necesario un procedimiento que permita hacer equivalentes los resultados obtenidos en diferentes formas o versiones de un mismo instrumento.

Para que dos formas de un instrumento de evaluación puedan ser equiparadas, se deben cubrir los siguientes requerimientos:

- Compartir las mismas características técnicas: estructura, especificaciones de reactivos, número de reactivos (longitud del instrumento) y un subconjunto de reactivos comunes (reactivos ancla), que en cantidad no deberá ser menor al 30% ni mayor al 50% de la totalidad de reactivos efectivos para calificar.

- Contar con una confiabilidad semejante.
- Los reactivos que constituyen el ancla deberán ubicarse en la misma posición relativa dentro de cada forma, y deberán quedar distribuidos a lo largo de todo el instrumento.
- La modalidad en la que se administren las formas deberá ser la misma para todos los aspirantes (por ejemplo, en lápiz y papel o en computadora).

Si el número de sustentantes es de al menos 100 en las distintas formas en que se llevará a cabo la equiparación, se utilizará el método de equiparación lineal para puntajes observados. Si el número de sustentantes es menor de 100 en alguna de las formas, se utilizará el método de equiparación de identidad (ver anexo técnico).

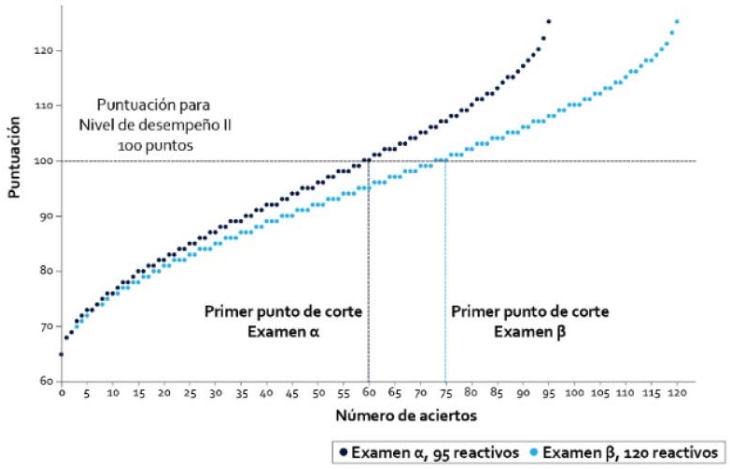
3.2 Escala utilizada para reportar los resultados

En cada plan de evaluación es indispensable definir la escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. En las escalas referidas a criterio cada calificación en la escala representa un nivel particular de desempeño referido a un estándar previamente definido en un campo de conocimiento o habilidad específicos.

El escalamiento que se llevará a cabo en este proceso de evaluación, permitirá construir una métrica común para todos los instrumentos que se administrarán. Consta de dos transformaciones, la primera denominada doble arcoseno, que permite estabilizar la magnitud de la precisión de las puntuaciones a lo largo de la escala; la segunda transformación es lineal y ubica el punto de corte del nivel de desempeño II en un mismo valor para todos los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos[1]).

Al utilizar esta escala, diferente a las escalas que se utilizan para reportar resultados de aprendizaje en el aula (de 5 a 10 o de 0% a 100%, donde el 6 o 60% de aciertos es aprobatorio), se evita que se realicen interpretaciones equivocadas de los resultados obtenidos en los exámenes, en virtud de que en los exámenes del spd cada calificación representa un nivel particular de desempeño respecto a un estándar previamente definido, el cual puede implicar un número de aciertos diferente en cada caso.

En la siguiente gráfica puede observarse el número de aciertos obtenidos en dos instrumentos de longitudes diferentes y con puntos de corte distintos que, a partir del escalamiento, es posible graficar en una misma escala, trasladando el primer punto de corte a 100 puntos, aun cuando en cada examen el punto de corte refiera a número de aciertos diferente. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



[1] Pueden encontrarse ligeras variaciones en este rango debido a que la escala es aplicable a múltiples instrumentos con características diversas, tales como la longitud, el tipo de instrumento y su nivel de precisión, diferencias entre los puntos de corte que atienden a las particularidades de los contenidos que se evalúan, entre otras. Por otra parte, para realizar el escalamiento, el sustentante debe, al menos, alcanzar un acierto en el examen; en caso contrario, se reportará como cero puntos y obtendrá NI.
En el caso de que el sustentante no presente alguno de los instrumentos de la evaluación, su resultado en ese instrumento será "NP: no presentó" y únicamente tendrá la devolución en aquellos instrumentos en los que haya participado y de los que se cuente con información, su resultado en el proceso de evaluación será "No idóneo" y no formará parte de los grupos de desempeño ni de las listas de prelación que publica la CNSPD a través de su página web.
Cuando el sustentante no presente ninguno de los instrumentos de la evaluación, su resultado en cada instrumento será "NP: no presentó", su resultado en el proceso de evaluación será "No se presentó a la evaluación" y no formará parte de los grupos de desempeño ni de las listas de prelación que publica la CNSPD a través de su página web.
Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

4. Resultado del proceso de evaluación: conformación de los grupos de desempeño

Dado que en cada instrumento se miden dominios diferentes y se atiende una lógica propia de diseño, construcción e incluso calificación, en ningún caso podrán sumarse el número de aciertos de cada examen para generar una puntuación global de todo el proceso de evaluación. Por esta razón, para determinar la idoneidad del resultado de cada aspirante se deberán integrar los resultados de todos los instrumentos sustentados, bajo el criterio de que:

> El sustentante con un resultado Idóneo en el proceso de evaluación será aquel que obtenga, al menos, el nivel de desempeño II (NII) en todos y cada uno de los instrumentos de evaluación que lo constituyen, según se define en los lineamientos del Concurso.

Cada sustentante conocerá su resultado integrado de todo el proceso de evaluación, así como sus resultados en cada uno de los exámenes que haya presentado.

Conformación de los grupos de desempeño

Con el conjunto de sustentantes que obtengan un resultado Idóneo en el proceso de evaluación, se conformarán grupos de desempeño en función de la combinación de resultados alcanzados del nivel de desempeño II o III (NII o NIII) en los instrumentos considerados en dicho proceso de evaluación. Los grupos de desempeño son el primer criterio de ordenamiento para la integración de las listas de prelación con los resultados idóneos que publica la CNSPD a través de su página web.

Todos los aspirantes sustentarán dos instrumentos de evaluación, por lo que los grupos de desempeño se conformarán de la manera siguiente:

> El **primer grupo de desempeño (A)**, se conformará con aquellos aspirantes que alcancen el nivel de desempeño III (NIII) en los dos exámenes involucrados en su proceso de evaluación.
>
> El **segundo grupo de desempeño (B)**, se conformará por los aspirantes que alcancen el nivel de desempeño II (NII) en uno de los exámenes y el nivel de desempeño III (NIII) en el otro examen.
>
> El **tercer grupo de desempeño (C)**, se conformará por los aspirantes que alcancen el nivel de desempeño II (NII) en los dos exámenes.

5. Integración de las listas de prelación

Las listas de prelación que son publicadas por la CNSPD se integrarán únicamente con los sustentantes que obtengan un resultado Idóneo en su proceso de evaluación y se ordenarán, en primer término, considerando los grupos de desempeño, iniciando con el grupo A, después el B y posteriormente el C, como se muestra en la siguiente tabla:

Tabla 2. Grupos de desempeño para los Concursos de oposición para la Promoción en Educación Básica y en Educación Media Superior

Distribución de exámenes y niveles de desempeño			Número de exámenes	Grupos y orden de ingreso en las listas de prelación
Nivel I	Nivel II	Nivel III		
-	-	2	2	1. Grupo A
-	1	1	2	2. Grupo B
-	2	-	2	3. Grupo C
1	-	1	2	No ingresan a la lista de prelación
1	1	-	2	
2	-	-	2	

Después, al interior de cada grupo de desempeño, se ordenará a los sustentantes considerando, primero, la puntuación obtenida en el examen de mayor relevancia o jerarquía de los dos instrumentos implicados en el proceso de evaluación y después la calificación obtenida en el otro instrumento. En todos los casos el ordenamiento de las listas de prelación será de mayor a menor puntuación.

El último criterio de ordenación está dado por las puntuaciones obtenidas por los sustentantes en contenidos de segundo nivel (por ejemplo, las áreas) de mayor importancia del examen de mayor jerarquía, después el puntaje obtenido en el área que le sigue de importancia, y así sucesivamente.

De manera excepcional, si después de llevar a cabo el proceso anterior para generar las listas de prelación se observan empates entre los resultados de algunos sustentantes, se recurrirá a la misma lógica de ordenamiento, tomando como referente el siguiente nivel de desagregación de los contenidos específicos que cuenten con la mayor cantidad de reactivos y que formen parte del examen que ha sido considerado como el de mayor relevancia. La jerarquía estará dada por el orden de la estructura del instrumento.

A continuación, se describe el orden de relevancia de los instrumentos y los contenidos al interior de cada uno de ellos, para Educación Básica y Educación Media Superior, que se utilizan en cada uno de los Concursos referidos en los párrafos anteriores, siendo de mayor relevancia el 1°, posteriormente el 2° y así sucesivamente.

Criterios específicos de ordenamiento para Educación Básica

En Educación Básica se consideran los Concursos de oposición para la promoción a cargos con funciones de dirección (Director, Subdirector y Coordinador de actividades), a cargos con funciones de supervisión (Jefe de sector, Jefe de enseñanza y Supervisor) y a las funciones de Asesoría Técnica Pedagógica (en Lenguaje oral y escrito, en Pensamiento matemático, en Educación especial y en Educación física).

A continuación, se indican las jerarquías de los instrumentos de evaluación para cada concurso, así como las relevancias de los contenidos específicos para ordenar los resultados e integrar las listas de prelación que son publicadas por la CNSPD.

- **Jerarquía de los instrumentos de evaluación para cargos con funciones de dirección: Director, Subdirector y Coordinador de actividades**

  1º Examen de conocimientos y habilidades para la práctica profesional

  2º Examen de habilidades intelectuales y responsabilidades ético profesionales

La jerarquía que se establece entre los dos exámenes se determina considerando que el primero de ellos expresa los conocimientos y las habilidades que debe poseer un directivo de Educación Básica, y que se consideran prioritarios e indispensables para el desempeño eficaz de la función, en este caso, el trabajo en el aula y las formas de organización y funcionamiento de la escuela para lograr que todos los alumnos aprendan.

El instrumento con el segundo nivel en la jerarquía evalúa conocimientos y habilidades que favorecen la mejora profesional, la comprensión de los principios filosóficos de la educación pública y el conocimiento del contexto social y cultural de la escuela, aspectos que complementan y fortalecen el ejercicio de la función.

La relevancia de los contenidos específicos de los dos instrumentos de evaluación se indica a continuación:

Instrumento de evaluación	Contenidos específicos	Funciones de Dirección / relevancia de los contenidos		
		Director	Subdirector	Coordinador de actividades
Examen de conocimientos y habilidades para la práctica profesional	La escuela y el trabajo en el aula	2°	1°	1°
	Gestión escolar	1°	2°	2°
Examen de habilidades intelectuales y responsabilidades ético profesionales	Mejora profesional para asegurar la calidad educativa	3°	3°	3°
	Principios para asegurar una educación de calidad	1°	1°	1°
	Influencia del contexto escolar	2°	2°	2°

En la función de Director es prioritaria la gestión escolar, entendida en sentido amplio, ya que no sólo se centra en la administración y uso de recursos, sino también en priorizar el aprendizaje de los alumnos y garantizar el funcionamiento eficaz de la escuela. Para el subdirector y el coordinador de actividades, su principal tarea radica en apoyar el desarrollo del currículo, el proceso de aprendizaje de los alumnos, así como el trabajo docente.

- **Jerarquía de los instrumentos de evaluación para cargos con funciones de supervisión: Jefe de sector, Jefe de enseñanza y Supervisor**

  1º Examen de conocimientos y habilidades para la práctica profesional

  2º Examen de habilidades intelectuales y responsabilidades ético profesionales

La jerarquía que se establece entre los dos exámenes se determina considerando que el primero de ellos expresa los conocimientos y las habilidades que debe poseer un supervisor escolar de Educación Básica, y que se consideran prioritarios e indispensables para el desempeño eficaz del cargo, en este caso, las funciones, la práctica profesional del personal, la asesoría pedagógica, la autonomía de gestión escolar y el vínculo con las autoridades educativas.

El instrumento con el segundo nivel en la jerarquía evalúa conocimientos y habilidades que favorecen la mejora profesional, la comprensión de los principios filosóficos de la educación pública y el conocimiento del contexto social y cultural de las escuelas y la forma de establecer relaciones de colaboración con las familias, las comunidades y otras instancias, aspectos que complementan y fortalecen el ejercicio de la función.

La relevancia de los contenidos específicos de los dos instrumentos de evaluación se indica a continuación:

Instrumento de evaluación	Contenidos específicos	Funciones de Supervisión / relevancia de los contenidos		
		Jefe de sector	Jefe de enseñanza	Supervisor
Examen de conocimientos y habilidades para la práctica profesional	Aspectos de la función	2°	1°	2°
	Gestión escolar	1°	2°	1°
Examen de habilidades intelectuales y responsabilidades ético profesionales	Mejora profesional para asegurar la calidad educativa	3°	3°	3°
	Principios para asegurar una educación de calidad	1°	1°	1°
	Influencia del contexto escolar	2°	2°	2°

En el caso del Jefe de sector y del Supervisor la prioridad radica en promover la autonomía de gestión escolar, garantizar el funcionamiento eficaz de las escuelas y zonas escolares a su cargo; mientras que para el jefe de enseñanza es prioritaria la asesoría académica, el desarrollo profesional de los docentes, la planeación didáctica, el seguimiento y la evaluación del aprendizaje.

- **Jerarquía de los instrumentos de evaluación para funciones de Asesoría Técnica Pedagógica (ATP)**

  1º Examen de conocimientos y habilidades para la práctica profesional

  2º Examen de habilidades intelectuales y responsabilidades ético profesionales

La jerarquía que se establece entre los dos exámenes se determina considerando que el primero de ellos expresa los conocimientos y las habilidades que debe poseer un Asesor Técnico Pedagógico de Educación Básica, y que se consideran prioritarios e indispensables para el desempeño eficaz de la función, en este caso, el aprendizaje infantil, la atención a la inclusión y la diversidad, las prácticas docentes, el currículo y la asesoría técnica pedagógica.

El instrumento con el segundo nivel en la jerarquía evalúa conocimientos y habilidades que favorecen la mejora profesional, la comprensión de los principios filosóficos de la educación pública y la capacidad para vincularse con la comunidad en que se ubica la escuela, con diferentes agentes educativos y con los integrantes de la zona escolar.

La relevancia de los contenidos específicos de los dos instrumentos de evaluación se indica a continuación:

Instrumento de evaluación	Contenidos específicos	Funciones de Asesoría Técnica Pedagógica / relevancia de los contenidos			
		Lenguaje oral y escrito	Pensamiento matemático	Educación especial	Educación física
Examen de conocimientos y habilidades para la práctica profesional	Dominio pedagógico	1°	1°	1°	1°
	Estrategias para la asesoría	2°	2°	2°	2°
Examen de habilidades intelectuales y responsabilidades ético profesionales	Mejora profesional para asegurar la calidad educativa	3°	3°	3°	3°
	Principios para asegurar una educación de calidad	1°	1°	1°	1°
	Influencia del contexto escolar	2°	2°	2°	2°

En el caso de los Asesores técnicos pedagógicos son prioritarios el dominio del Plan y programas de estudio: enfoques de enseñanza, estrategias didácticas, uso de materiales educativos, planeación y evaluación de la enseñanza y el aprendizaje.

Criterios específicos de ordenamiento para Educación Media Superior

Para Educación Media Superior se consideran los Concursos de oposición para la promoción a cargos con funciones de dirección (Director, Jefe de departamento académico, Jefe de departamento administrativo, Subdirector académico y Subdirector administrativo) y a cargos con funciones de supervisión (Supervisor). A continuación, se indican las jerarquías de los instrumentos de evaluación para cada concurso, así como las relevancias de los contenidos específicos para ordenar los resultados e integrar las listas de prelación que son publicadas por la CNSPD.

Concursos de oposición para cargos con funciones de dirección

- **Jerarquía de los instrumentos de evaluación para Director**

  1º Plan de mejora de la función

  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Normativa de la función directiva	1°	La correcta aplicación de la normativa de la función directiva es fundamental para la operación del plantel, por lo que se subraya la importancia de que los directivos cuenten con determinadas competencias para el cumplimiento de su función.
Gestión institucional	2°	La gestión eficiente de un plantel requiere de cualidades, habilidades y actitudes necesarias para generar el liderazgo que promueva la óptima operatividad del plantel, en apego a los marcos normativos que regulan el funcionamiento del mismo.

El instrumento de evaluación **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

- **Jerarquía de los instrumentos de evaluación para Subdirector académico**

  1º Plan de mejora de la función

  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Conocimientos de la función	1°	El conocimiento de los lineamientos de operación del plantel y de los procesos de gestión académica son fundamentales en la implementación del trabajo colaborativo y colegiado para alcanzar el perfil de egreso de la EMS.
Responsabilidades y ética del subdirector	2°	El conocimiento de los principios que rigen su función da sustento a las acciones de intervención que realiza, de acuerdo con la normatividad aplicable.

El instrumento **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

- **Jerarquía de los instrumentos de evaluación para Subdirector administrativo**

  1º Plan de mejora de la función

  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Gestión Administrativa	1°	La gestión de los recursos humanos, financieros y materiales, en apego al marco normativo, es la función primordial del Subdirector administrativo, para la operación y buen funcionamiento de un plantel.
Planeación estratégica	2°	La organización y desarrollo de las estrategias de administración para el cumplimiento de metas y objetivos bajo criterios de transparencia y rendición de cuentas, permite elevar la calidad del servicio en el plantel.

El instrumento **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

- **Jerarquía de los instrumentos de evaluación para Jefe de departamento académico**

  1º Plan de mejora de la función

  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Funcionamiento del departamento académico	1°	El conocimiento de las funciones que realiza el Jefe de departamento académico es fundamental para implementar acciones eficaces que permitan elevar los indicadores académicos del plantel.
Mejora de la función del Jefe de departamento académico	2°	El conocimiento de los lineamientos normativos y éticos que regulan la función contribuye a orientar la mejora de los procesos académicos del plantel.

El instrumento **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

- **Jerarquía de los instrumentos de evaluación para Jefe de departamento administrativo**
  1º Plan de mejora de la función
  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Marco normativo para el funcionamiento del plantel	1°	El conocimiento de la normativa que regula la función del jefe de departamento administrativo es fundamental para el funcionamiento del plantel; sin embargo, para su correcta aplicación se requiere además de contar con determinados conocimientos para la operación administrativa.
Función administrativa para la operación del plantel	2°	La organización de las actividades asociadas a la función del Jefe de departamento administrativo es necesaria para la optimización de los recursos humanos, financieros y materiales con el fin de alcanzar los propósitos educativos del plantel.
Desarrollo de la función administrativa	3°	El desarrollo de la función requiere del conocimiento de las responsabilidades éticas y de rendición de cuentas para la asignación adecuada de los recursos del plantel.

El instrumento **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

Concursos de oposición para cargos con funciones de supervisión

- **Jerarquía de los instrumentos de evaluación para Supervisor**
  1º Plan de mejora de la función
  2º Examen de conocimientos

Para el **Examen de conocimientos** se considerará el orden de relevancia que se indica a continuación:

Contenidos específicos	Relevancia	Argumentos
Desarrollo profesional	1°	El conocimiento de las acciones y mecanismos que debe implementar un Supervisor es fundamental para fortalecer el proceso educativo de los planteles de acuerdo con su contexto.
Ética y normatividad de las funciones del Supervisor	2°	El conocimiento del marco ético y normativo que regula y orienta la función de Supervisor contribuye a garantizar la calidad de la Educación Media Superior.

El instrumento **Plan de mejora de la función** no ingresa al último criterio de ordenamiento debido a que se evalúa a través de una rúbrica y no se contemplan niveles de desagregación de sus contenidos específicos, organizados a través de tareas evaluativas.

Finalmente, cualquier situación no prevista en los presentes criterios técnicos será analizada por la Junta de Gobierno del INEE para emitir una determinación, según corresponda con el marco normativo vigente.

Sobre la integralidad de la evaluación para emitir la calificación

Dado que los presentes criterios técnicos se han definido con el objetivo de aportar evidencia para la validez de las inferencias que se desean obtener a partir de los datos recopilados, es fundamental señalar que, ***en ningún caso, se puede considerar solamente un instrumento para integrar el resultado de los sustentantes conforme al diseño de la evaluación***. Lo anterior debido a que la evidencia empírica que resulte del análisis psicométrico de los instrumentos debe mostrar que, una vez que éstos fueron aplicados, cumplen con la calidad técnica indispensable para emitir un resultado toda vez que la integración de los resultados de la evaluación debe permitir establecer inferencias válidas sobre el desempeño y competencias de los sustentantes evaluados.

Anexo técnico

El propósito de este anexo es detallar los aspectos técnicos específicos de los distintos procedimientos que se han enunciado en el cuerpo del documento, así como de brindar mayores elementos para su entendimiento y fundamento metodológico.

Protocolo de calificación por jueces para las rúbricas

A continuación, se presenta un protocolo que recupera propuestas sistemáticas de la literatura especializada (Jonsson y Svingby, 2007; Rezaei y Lovorn, 2010; Stemler y Tsai, 2008; Stellmack, et. al, 2009).

1. Se reciben las evidencias de evaluación de los sustentantes, mismas que deben cumplir con las características solicitadas por la autoridad educativa.

2. Se da a conocer a los jueces la rúbrica de calificación y se les capacita para su uso.

3. Las evidencias de los sustentantes son asignadas de manera aleatoria a los jueces, por ejemplo se pueden considerar *redes no dirigidas*; intuitivamente, una red no dirigida puede pensarse como aquella en la que las conexiones entre los nodos siempre son simétricas (si A está conectado con B, entonces B está conectado con A y sucesivamente con los *n* número de jueces conectados entre sí), este tipo de asignación al azar permite contar con indicadores iniciales de cuando un juez está siendo reiteradamente "estricto" o reiteradamente "laxo" en la calificación, lo cual ayudará a saber si es necesario volver a capacitar a alguno de los jueces y permitirá obtener datos de consistencia inter-juez.

4. Cada juez califica de manera individual las evidencias sin conocer la identidad de los sustentantes o cualquier otro dato que pudiera alterar la imparcialidad de la decisión del juez.

5. Los jueces emiten la calificación de cada sustentante, seleccionando la categoría de ejecución que consideren debe recibir el sustentante para cada uno de los aspectos a evaluar que constituyen la rúbrica, esto en una escala ordinal (por ejemplo: de 0 a 3, de 0 a 4, de 1 a 6, etc.), lo pueden hacer en un formato impreso o electrónico a fin de conservar dichas evidencias.

6. Si existen discrepancias entre los dos jueces en cuanto a la asignación de categorías en algunos aspectos a evaluar se deben tomar decisiones al respecto. A continuación, se muestran orientaciones para esta toma de decisiones:

a) Cuando la discrepancia corresponde a categorías de ejecución contiguas (por ejemplo: 1-2) se asigna la categoría superior. Esto permite favorecer al sustentante ante dicho desacuerdo entre los jueces.

b) Cuando son categorías no contiguas de la rúbrica:

- Si existe solamente una categoría en medio de las decisiones de los jueces (por ejemplo: 1-3), se asigna al sustentante la categoría intermedia. No se deben promediar los valores asignados a las categorías.
- Si existe más de una categoría en medio de las decisiones de los jueces (por ejemplo: 1-4), se debe solicitar a los jueces que verifiquen si no hubo un error al momento de plasmar su decisión. En caso de no haber ajustes por este motivo, se requiere la intervención de un tercer juez, quien debe asignar la categoría de ejecución para cada uno de los aspectos a evaluar; la categoría definitiva que se asigna al sustentante en cada aspecto a evaluar debe considerar las decisiones de los dos jueces que den mayor puntaje total al sustentante, si existe discrepancia en algún aspecto a evaluar se asigna la categoría superior, a fin de favorecer al sustentante ante dicho desacuerdo entre los jueces.

7. Los jueces firman la evidencia con las asignaciones de categorías definitivas en cada aspecto a evaluar.

8. La calificación global del sustentante se determina de la siguiente forma:

a. Se identifica la categoría asignada al sustentante en cada aspecto a evaluar.

b. Se identifica el valor asignado a cada categoría de la rúbrica.

c. La suma de los valores es el resultado de la calificación.

9. Las asignaciones de categorías en cada aspecto a evaluar para emitir la calificación global definitiva del sustentante son plasmadas en algún formato impreso o electrónico, con la debida firma, autógrafa o electrónica de los jueces, a fin de que queden resguardadas como evidencia del acuerdo de la calificación definitiva del proceso de jueceo.

Métodos para establecer puntos de corte y niveles de desempeño

Método de Angoff

El método de Angoff está basado en los juicios de los expertos sobre los reactivos y contenidos que se evalúan a través de exámenes. De manera general, el método considera que el punto de corte se define a partir de la ejecución promedio de un sustentante hipotético que cuenta con los conocimientos, habilidades o destrezas que se consideran indispensables para la realización de una tarea en particular; los jueces estiman, para cada pregunta, cuál es la probabilidad de que dicho sustentante acierte o responda correctamente.

Procedimiento

Primero se juzgan algunas preguntas, con tiempo suficiente para explicar las razones de las respuestas al grupo de expertos y que les permite homologar criterios y familiarizarse con la metodología.

Posteriormente, se le solicita a cada juez que estime la probabilidad mínima de que un sustentante conteste correctamente un reactivo, el que le sigue y así hasta concluir con la totalidad de los reactivos, posteriormente se calcula el puntaje esperado (*raw score*: la suma de estas probabilidades multiplicadas por uno para el caso de reactivos –toda vez que cada reactivo vale un punto–; o bien, la suma de estas probabilidades multiplicadas por el valor máximo posible de las categorías de la rúbrica). Las decisiones de los jueces se promedian obteniendo el punto de corte. La decisión del conjunto de jueces pasa por una primera ronda para valorar sus puntos de vista en plenaria y puede modificarse la decisión hasta llegar a un acuerdo en común.

Método de Beuk

En 1981, Cess H. Beuk propuso un método para establecer estándares de desempeño, el cual busca equilibrar los juicios de expertos basados solamente en las características de los instrumentos de evaluación, lo que mide y su nivel de complejidad, con los juicios que surgen del análisis de resultados de los sustentantes una vez que un instrumento de evaluación es administrado.

Procedimiento

En el cuerpo del documento se señalaron tres fases para el establecimiento de los puntos de corte. Para completar la tercera fase, es necesario recolectar con antelación las respuestas a dos preguntas dirigidas a los integrantes de los distintos comités académicos especializados involucrados en el diseño de las evaluaciones y en otras fases del desarrollo del instrumento. Las dos preguntas son:

a) ¿Cuál es el mínimo nivel de conocimientos o habilidades que un sustentante debe tener para aprobar el instrumento de evaluación? (expresado como porcentaje de aciertos de todo el instrumento, k).

b) ¿Cuál es la tasa de aprobación de sustentantes que los jueces estiman que aprueben el instrumento? (expresado como porcentaje, v).

Para que los resultados de la metodología a implementar sean estables e integren diferentes enfoques que contribuyan a la diversidad cultural, se deberán recolectar las respuestas de, al menos, 30 especialistas integrantes de los diferentes comités académicos que hayan participado en el diseño y desarrollo de los instrumentos.

Adicionalmente, se debe contar con la distribución de los sustentantes para cada posible punto de corte, con la finalidad de hacer converger el juicio de los expertos con la evidencia empírica.

Los pasos a seguir son los siguientes:

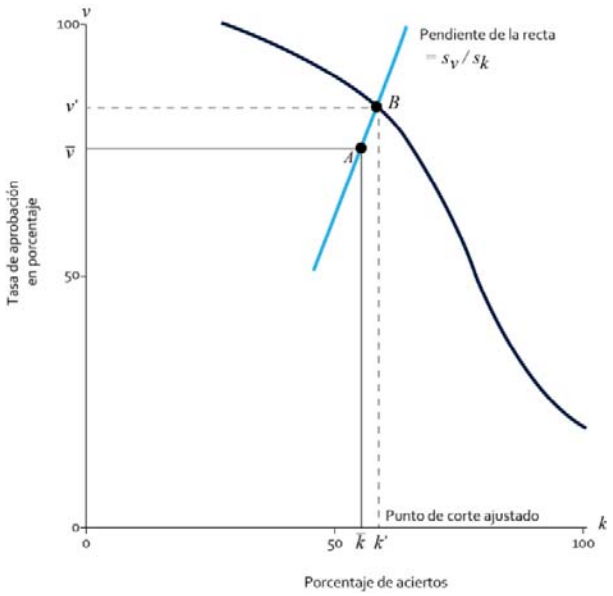
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas ($\bar{k}$, $\bar{v}$), (ver siguiente figura).

2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

3. Se calcula la desviación estándar de k y v (s_k y s_v).

4. A partir del punto A se proyecta una recta con pendiente s_v/s_k hasta la curva de distribución empírica (del paso 2). El punto de intersección entre la recta y la curva de distribución es el punto B. La recta se define como: $v = (s_v/s_k)(k - \bar{k}) + \bar{v}$.

El punto B, el cual tiene coordenadas (k', v'), representa los valores ya ajustados, por lo que k' corresponderá al punto de corte del estándar de desempeño. El método asume que el grado en que los expertos están de acuerdo es proporcional a la importancia relativa que los expertos dan a las dos preguntas, de ahí que se utilice una línea recta con pendiente s_v/s_k.



Escalamiento de las puntuaciones

El escalamiento (Wilson, 2005) se llevará a cabo a partir de las puntuaciones crudas de los sustentantes, y se obtendrá una métrica común para los instrumentos de evaluación, que va de 60 a 170 puntos, aproximadamente, ubicando el primer punto de corte (nivel de desempeño NII) para los instrumentos en los **100 puntos**. El escalamiento consta de dos transformaciones:

a) Transformación doble arcoseno que estabiliza la magnitud de la precisión que se tiene para cada punto de la escala.

b) Transformación lineal que ubica el primer punto de corte en 100 unidades y define el número de distintos puntos en la escala (el rango de las puntuaciones) con base en la confiabilidad del instrumento, por lo que, a mayor confiabilidad, habrá más puntos en la escala (Shun-Wen Chang, 2006).

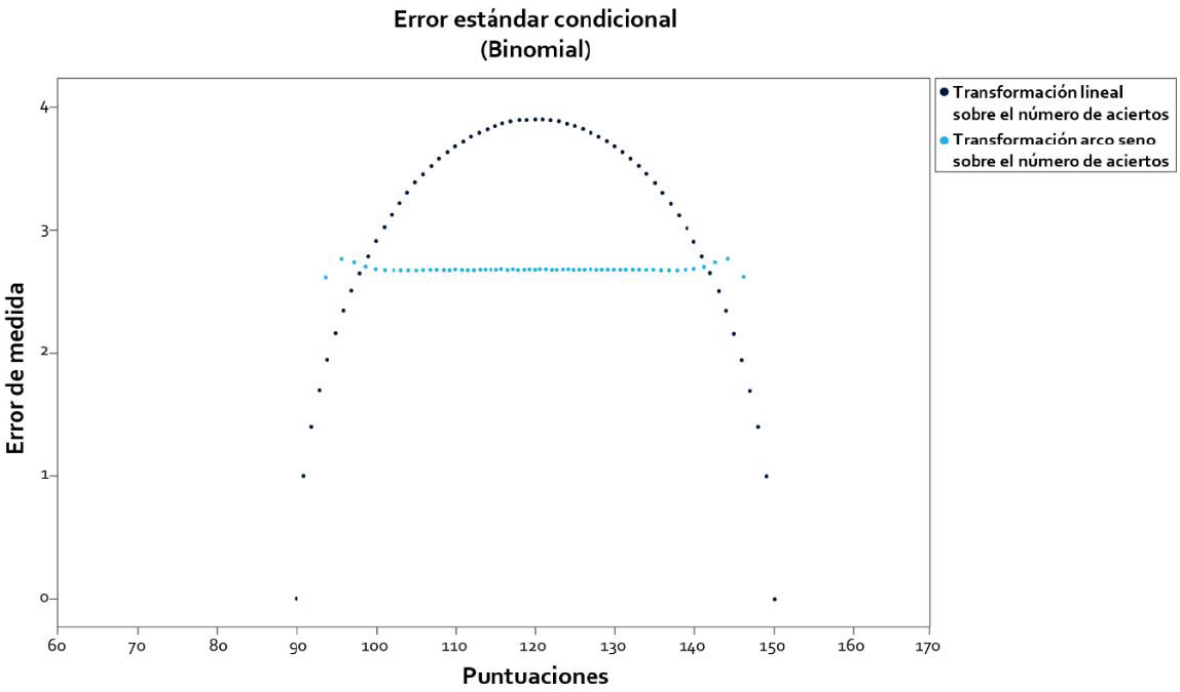
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

Finalmente, es importante destacar que para que se lleve a cabo el escalamiento, el sustentante debió alcanzar, al menos, un acierto en el instrumento de evaluación en cuestión. De no ser así, se reportará como cero y el resultado será NI.

Procedimiento para la transformación doble arcoseno

En los casos de los exámenes de opción múltiple, deberá calcularse el número de respuestas correctas que haya obtenido cada sustentante en el instrumento de evaluación. Los reactivos se calificarán como correctos o incorrectos de acuerdo con la clave de respuesta correspondiente. Si un sustentante no contesta un reactivo o si selecciona más de una alternativa de respuesta para un mismo reactivo, se calificará como incorrecto. Cuando los instrumentos de evaluación sean calificados con rúbricas, deberá utilizarse el mismo procedimiento para asignar puntuaciones a los sustentantes considerando que *K* sea la máxima puntuación que se pueda obtener en el instrumento de evaluación.

Cuando se aplica la transformación doble arcoseno sobre el número de aciertos obtenido en el instrumento de evaluación, el error estándar condicional de medición de las puntuaciones obtenidas se estabiliza, es decir, es muy similar, pero no igual, a lo largo de la distribución de dichas puntuaciones, con excepción de los valores extremos, a diferencia de si se aplica una transformación lineal, tal y como se observa en la siguiente gráfica (Won-Chan, Brennan y Kolen, 2000).



Para estabilizar la varianza de los errores estándar condicionales de medición a lo largo de la escala y por tanto medir con similar precisión la mayoría de los puntajes de la escala, se utilizará la función *c:*

$$c(k_i) = \frac{1}{2}\left\{\operatorname{arcsen}\sqrt{\frac{k_i}{K+1}} + \operatorname{arcsen}\sqrt{\frac{k_i+1}{K+1}}\right\} \quad (1)$$

Donde:

i se refiere a un sustentante

k_i es el número de respuestas correctas que el sustentante *i* obtuvo en el instrumento de evaluación

K es el número de reactivos del instrumento de evaluación

Procedimiento para la transformación lineal

Una vez que se aplica la transformación doble arcoseno que estabiliza la magnitud de la precisión que se tiene para cada punto de la escala, se procede a aplicar la transformación lineal que ubica el primer punto de corte en 100 unidades.

La puntuación mínima aceptable que los sustentantes deben tener para ubicarse en el nivel de desempeño NII en los instrumentos de evaluación, se situará en el valor 100. Para determinarla se empleará la siguiente ecuación:

$$P_i = A * c(k_i) + B \quad (2)$$

Donde $A = \frac{Q}{[c(K)-c(0)]}$, $B = 100 - A * c(PC1)$, *Q* es la longitud de la escala, *c(K)* es la función *c* evaluada en *K*, *c(0)* es la misma función *c* evaluada en cero y PC1 es el primer punto de corte (en número de aciertos) que se definió para establecer los niveles de desempeño y que corresponde al mínimo número de aciertos que debe tener un sustentante para ubicarlo en el nivel de desempeño NII.

El valor de *Q* dependerá de la confiabilidad del instrumento. Para confiabilidades iguales o mayores a 0.90, *Q* tomará el valor 80 y, si es menor a 0.90 tomará el valor 60 (Kolen y Brennan, 2014). Lo anterior implica que los extremos de la escala pueden tener ligeras fluctuaciones.

Por último, las puntuaciones P_i deben redondearse al entero más próximo, utilizando el criterio de que puntuaciones con cinco décimas o más, suben al siguiente entero.

Cálculo de las puntuaciones de los contenidos específicos de primer nivel en los instrumentos de evaluación

Para calcular las puntaciones del sustentante (i) en los contenidos específicos del primer nivel, se utilizará la puntuación ya calculada para el examen (P_i), el número de aciertos de todo el instrumento de evaluación (k_i), y el número de aciertos de cada uno de los contenidos específicos que conforman el instrumento (k_{Aji}). Las puntuaciones de los contenidos específicos (P_{Aji}) estarán expresadas en números enteros y su suma deberá ser igual a la puntuación total del instrumento (P_i).

Si el instrumento de evaluación está conformado por dos contenidos específicos, primero se calculará la puntuación del contenido específico que tenga la mayor relevancia (P_{A1i}), de acuerdo con la jerarquización de los contenidos de primer nivel (que se indicaron en el apartado de la definición de las listas de prelación), mediante la ecuación:

$$\boldsymbol{P_{A1i} = P_i * \frac{k_{A1i}}{k_i}} \quad (3)$$

El resultado se redondeará al entero inmediato anterior con el criterio de que puntuaciones con cinco décimas o más suben al siguiente entero. La otra puntuación del contenido específico del primer nivel (P_{A2i}) se calculará como:

$$\boldsymbol{P_{A2i} = P_i - P_{A1i}} \quad (4)$$

Para los instrumentos de evaluación con más de dos contenidos específicos, se calculará la puntuación de cada uno siguiendo el mismo procedimiento empleando la ecuación (3) para los primeros. La puntuación del último contenido específico, que tiene una menor prioridad, se calculará por sustracción como complemento de la puntuación del instrumento de evaluación, el resultado se redondeará al entero positivo más próximo. De esta manera, si el instrumento consta de *j* contenidos específicos, la puntuación del *j-ésimo* contenido específico será:

$$\boldsymbol{P_{Aji} = P_i - \sum_{k=1}^{j-1} P_{Aki}} \quad (5)$$

En los casos donde el número de aciertos de un conjunto de contenidos específicos del instrumento sea cero, no se utilizará la fórmula (3) debido a que no está definido el valor de un cociente en donde el denominador tome el valor de cero. En este caso, el puntaje deberá registrarse como cero.

Procedimiento para el error estándar condicional. Método delta

Dado que el error estándar de medición se calcula a partir de la desviación estándar de las puntuaciones y su correspondiente confiabilidad, dicho error es un 'error promedio' de todo el instrumento. Por lo anterior, se debe implementar el cálculo del error estándar condicional de medición (CSEM), que permite evaluar el error estándar de medición (SEM) para puntuaciones específicas, por ejemplo, los puntos de corte.

Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta, (Muñiz, 2003), que calcula los errores estándar de medición condicionales. Para incluir la confiabilidad del instrumento de medición se usa un modelo de error binomial, para el cálculo del error estándar condicional de medición será:

$$\boldsymbol{\sigma(X) = \sqrt{\frac{1-\alpha}{1-KR21}\left[\frac{X(n-X)}{n-1}\right]}}$$

Donde:

X es una variable aleatoria asociada a los puntajes

n es el número de reactivos del instrumento

KR21 es el coeficiente de Kuder-Richardson.

α es el coeficiente de confiabilidad de Cronbach, KR-20 (Thompson, 2003):

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n}\sigma_J^2}{\sigma_X^2}\right)$$

$\sum_{j=1}^{n}\sigma_J^2$ = suma de las varianzas de los *n* reactivos

σ_X^2 = varianza de las puntuaciones en el instrumento

Para calcular el error estándar condicional de medición de la transformación P_i, se emplea el Método delta, el cual establece que si $P_i = g(X)$, entonces un valor aproximado de la varianza de $g(X)$ está dado por:

$$\sigma^2(P_i) \doteq \left(\frac{dg(X)}{dX}\right)^2 \sigma^2(X)$$

De ahí que:

$$\sigma(P_i) \doteq \frac{dg(x)}{dx}\,\sigma(x)$$

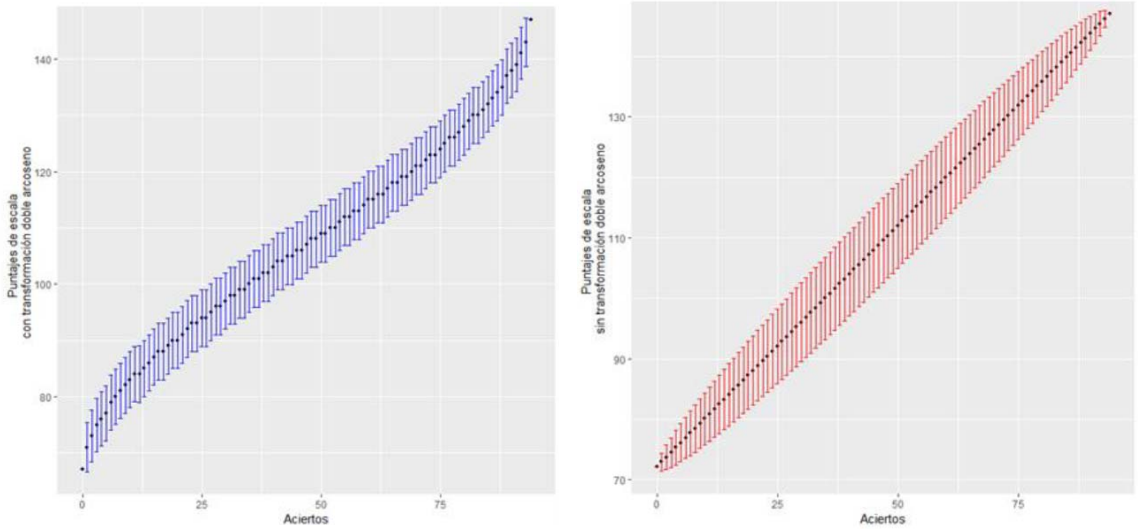
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

$$\sigma(P_i) \doteq \frac{A}{2}\left[\frac{1}{2(k+1)\left(\sqrt{\frac{x}{k+1}}\right)\left(\sqrt{1-\frac{x}{k+1}}\right)} + \frac{1}{2(k+1)\left(\sqrt{\frac{x+1}{k+1}}\right)\left(\sqrt{1-\frac{x+1}{k+1}}\right)}\right]\sigma(x)$$

Donde $\sigma(x)$ es el error estándar de medida de las puntuaciones crudas y $\sigma(P_i)$ el error estándar condicional de medición, de la transformación P_i, que ya incorpora la confiabilidad.

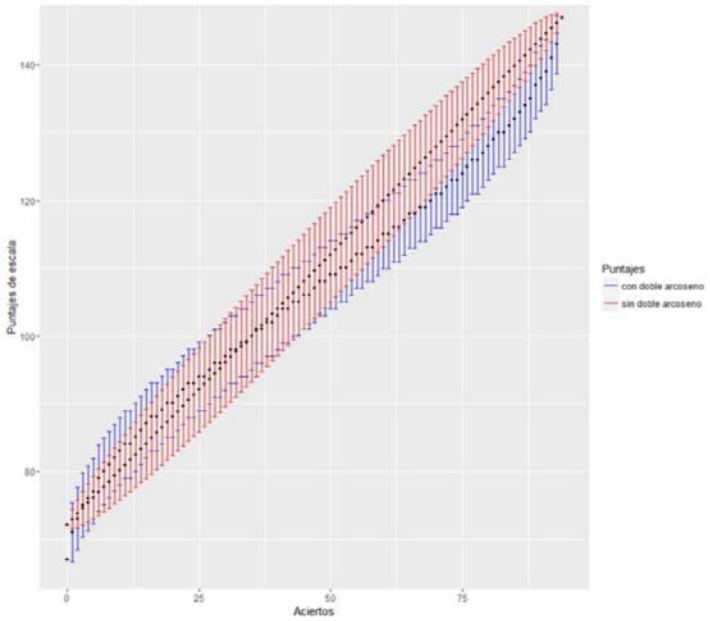
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 2012; American College Testing, 2013; 2014a; 2014b).

En las siguientes gráficas se muestran los intervalos de confianza (al 95% de confianza) de los puntajes de la escala cuando se aplica la transformación doble arcoseno (gráfica del lado izquierdo) y cuando no se aplica (gráfica del lado derecho).



Se observa que al aplicar la transformación doble arcoseno se mide con similar precisión la mayoría de los puntajes de la escala, a diferencia de cuando no se aplica dicha transformación, además de que en el punto de corte para alcanzar el nivel de desempeño NII (100 puntos) el error es menor cuando se aplica la transformación.

Esto es más claro si se observan ambas gráficas en el mismo cuadrante, como en la siguiente imagen.



El dato obtenido del error estándar condicional deberá reportarse en la misma escala en que se comunican las calificaciones de los sustentantes e incorporarse en el informe o manual técnico del instrumento (estándar 2.13 de los Estándares para las Pruebas Educativas y Psicológicas de la *American Educational Research Association* et. al., 2014). Asimismo, esto permite atender al estándar 2.14 de los Estándares para las Pruebas Educativas y Psicológicas de la *American Educational Research Association* et. al., 2014, el cual establece que cuando se especifican puntos de corte para selección o clasificación, los errores estándar deben ser reportados en la vecindad de cada punto de corte en dicho informe o manual técnico.

Proceso para la equiparación de instrumentos de evaluación

El procedimiento que permite hacer equivalentes los resultados obtenidos en diferentes formas o versiones de un mismo instrumento es una equiparación. La que aquí se plantea considera dos estrategias: a) si el número de sustentantes es de al menos 100 en ambas formas, se utilizará el método de equiparación lineal de Levine para puntajes observados; o bien, b) si el número de sustentantes es menor de 100 en alguna de las formas, se utilizará el método de equiparación de identidad (*identity equating*). A continuación, se detallan los procedimientos.

Método de equiparación lineal de Levine

La equiparación de las formas de un instrumento deberá realizarse utilizando el método de equiparación lineal de Levine (Kolen y Brennan, 2014), para puntajes observados bajo un diseño de grupos no equivalentes con reactivos comunes. Dicho diseño es uno de los más utilizados en la práctica. En cada muestra de sujetos se administra solamente una forma de la prueba, con la peculiaridad de que en ambas muestras se administra un conjunto de reactivos en común llamado ancla, que permite establecer la equivalencia entre las formas a equiparar.

Cualquiera de los métodos de equiparación de puntajes que se construya involucra dos poblaciones diferentes.

Sin embargo, una función de equiparación de puntajes se define sobre una población única. Por lo tanto, las poblaciones 1 y 2 que corresponden a las poblaciones donde se aplicó la forma nueva y antigua, deben ser combinadas para obtener una población única a fin de definir una relación de equiparación.

Esta única población se conoce como población sintética, en la cual se le asignan pesos w_1 y w_2 a las poblaciones 1 y 2, respectivamente, esto es, $\mathbf{w_1 + w_2 = 1}$ y $\mathbf{w_1, w_2 \geq 0}$. Para este proceso se utilizará

$$\boldsymbol{w_1 = \frac{N_1}{N_1 + N_2}}$$

y

$$\boldsymbol{w_2 = \frac{N_2}{N_1 + N_2}}$$

Donde N_1 corresponde al tamaño de la población 1 y N_2 corresponde al tamaño de la población 2.

Los puntajes de la forma nueva, aplicada a la población 1, serán denotados por X; los puntajes de la forma antigua, aplicada a la población 2, serán denotados por Y.

Los puntajes comunes están identificados por V y se dice que los reactivos comunes corresponden a un anclaje interno cuando V se utiliza para calcular los puntajes totales de ambas poblaciones.

Usando el concepto de población sintética, la relación lineal de equiparación de puntajes para el diseño de grupos no equivalentes con reactivos comunes se escribe de la siguiente forma:

$$\boldsymbol{l_{Y_s}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y)}$$

Donde s denota la población sintética y

$$\boldsymbol{\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)]}$$

$$\boldsymbol{\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)]}$$

$$\boldsymbol{\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_1^2[\mu_1(V) - \mu_2(V)]^2}$$

$$\boldsymbol{\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1w_2\gamma_2^2[\mu_1(V) - \mu_2(V)]^2}$$

Donde los subíndices 1 y 2 se refieren a las poblaciones 1 y 2 respectivamente.

$$\boldsymbol{\gamma_1 = \frac{\sigma_1(X,V)}{\sigma_1^2(V)}}$$

y

$$\boldsymbol{\gamma_2 = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}}$$

Específicamente, para el método de Levine para puntajes observados bajo un diseño de grupos no equivalentes con reactivos comunes, las γ´s se expresan de la siguiente manera:

$$\boldsymbol{\gamma_1 = \frac{\sigma_1^2(X)}{\sigma_1(X,V)}}$$

$$\boldsymbol{\gamma_2 = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)}}$$

Para aplicar este método basta con reemplazar estos coeficientes en las ecuaciones lineales antes descritas. Por su parte, Kolen y Brennan proveen justificaciones para usar esta aproximación.

Es importante señalar que para los puntajes que se les aplique la equiparación $\boldsymbol{x_e = b_1 x + b_0}$, con b_1 como pendiente y b_0 como ordenada al origen, el procedimiento es análogo al descrito en la sección "Procedimiento para el error estándar condicional. Método delta", y el error estándar condicional de medición para la transformación $\boldsymbol{P_{i_e} = A * c(x_e) + B}$, que ya incorpora la confiabilidad, está dado por:

$$\boldsymbol{\sigma(P_{i_e}) \doteq \frac{A}{2}\left[\frac{b1}{2(k+1)\left(\sqrt{\frac{x_e}{k+1}}\right)\left(\sqrt{1-\frac{x_e}{k+1}}\right)}+\frac{b1}{2(k+1)\left(\sqrt{\frac{x_e+1}{k+1}}\right)\left(\sqrt{1-\frac{x_e+1}{k+1}}\right)}\right]\sigma(x_e)}$$

Donde x_e son las puntuaciones equiparadas, las cuales son una transformación de las puntuaciones crudas, por lo que el error estándar de medida de dicha transformación se define como:

$$\boldsymbol{\sigma(x_e) = b_1 * \sigma(x)}$$

Método de equiparación de identidad (*identity equating*)

La equiparación de identidad es la más simple, toda vez que no hace ningún ajuste a la puntuación "*x*" en la escala de la forma *X* al momento de convertirla en la puntuación equiparada "y" en la escala de la forma *Y*.

Es decir, dichas puntuaciones son consideradas equiparadas cuando tienen el mismo valor, por lo que las coordenadas de la línea de equiparación de identidad están definidas simplemente como *x*=*y* (Holland y Strawderman, 2011).

Referencias

American College Testing. (2013). *ACT Plan Assessments Technical Manual*, Iowa City, IA: Author.

American College Testing. (2014a) *ACT Assessments Technical Manual*, Iowa City, IA: Author.

American College Testing. (2014b) *ACT QualityCore Assessments Technical Manual*, Iowa City, IA: Author.

AERA, APA y NCM. (2014). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Angoff, W. (1971). *Scales, norms, and equivalent scores*. En R. Thorndike (Ed.), Educational Measurement (2° ed.). Washington, DC: American Council on Education.

Beuk, C. H. (1984). A Method for Reaching a Compromise between Absolute and Relative Standards in Examinations. *Journal of Educational Measurement*, 21 (2) pp. 147-152.

Brennan, R. L. (2012). Scaling PARCC Assessments: Some considerations and a synthetic data example.

Cook, D. y Beckman, T. (2006). *Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application*. *The American Journal of Medicine 119*, 166.e7-166.e16

Downing, S. (2004). Reliability: On the reproducibility of assessment data. Med Educ 38(9) pp.1006-1012.

Holland, P. y Strawderman, W. (2011). How to average equating functions, if you must. En Von Davier, A. (Ed.). *Statistical models for test equating, scaling, and linking* (pp. 89–107). Nueva York, NY: Springer.

Jonsson, A. y Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2 pp.130-144.

Kendall, M. y Stuart, A. (1977). *The advanced theory of statistics, Vol. 1: Distribution theory*. 4ª Ed. New York, NY: MacMillan.

Kolen, M. y Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices*. 3ª Ed. New York, NY: Springer-Verlag.

Masters, G. (1982). *A Rasch model for Partial Credit Scoring*. Psychometrika. 47(2) pp 149-174.

Muñiz, J. (2003). *Teoría clásica de los test.* Madrid, España: Ediciones pirámide.

Muraki, E. (1999*). Stepwise Analysis of Differential Item Functioning Based on Multiple-Group Partial Credit Model.* Journal of Educational Measurement.

OECD (2002). *PISA 2000 Technical Report.* Paris, Francia: OECD Publishing.

OECD (2005). *PISA 2003 Technical Report*. Paris, Francia: OECD Publishing.

OECD (2009). *PISA 2006 Technical Report*. Paris, Francia: OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. Paris, Francia: OECD Publishing.

Rezaei, A. y Lovorn, M. (2010). *Reliability and validity of rubrics for assessment through writing. Assessing Writing* 15 (1) 18-39.

Chang, S. (2006). *Methods in Scaling the Basic Competence Test, Educational and Psychological Measurement,* 907-927

Stellmack, M., Konheim-Kalkstein, Y., Manor, J., Massey, A. y Schmitz, J. (2009). An assessment of reliability and validity of a rubric for APA-style introductions, *Teaching of Psychology,* 36 pp. 102-107.

Stemler, E. y Tsai, J. (2008). 3 Best Practices in Interrater Reliability Three Common Approaches*.* En *Best practices in quantitative methods.* pp. 29-49. California, E.U.A.: SAGE Publications, Inc.

Thompson, B. (ed.). (2003). *Score reliability. Contemporary thinking on reliability issues*. California, E.U.A.: SAGE Publications, Inc.

Wilson, M. (2005). *Constructing measures. An ítem response modeling approach*. Mahwah, Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

Won-Chan, L., Brennan, R. y Kolen, M. (2000). Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study. *Journal of Educational Measurement*, 37 (1) pp. 1-20.

Wu, M. y Adams, R. (2007). *Applying the Rasch Model to Psycho-social measurement. A practical approach.* Melbourne, Australia: Educational measurement solutions.

TRANSITORIOS

Primero. Los presentes Criterios entrarán en vigor al día siguiente de su publicación en el Diario Oficial de la Federación.

Segundo. Los presentes Criterios, de conformidad con los artículos 40 y 48 de la Ley del Instituto Nacional para la Evaluación de la Educación, deberán hacerse del conocimiento público a través de la página de Internet del Instituto www.inee.edu.mx

Tercero. Se instruye a la Dirección General de Asuntos Jurídicos para que realice las gestiones necesarias a efecto de que los presentes Criterios se publiquen en el Diario Oficial de la Federación.

Ciudad de México, a trece de marzo de dos mil diecinueve. - Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Cuarta Sesión Extraordinaria de dos mil diecinueve, celebrada el trece de marzo de dos mil diecinueve. - Acuerdo número SEJG/04-19/04,R. La Consejera Presidenta, **Teresa Bracho González**.- Los Consejeros: **Bernardo Hugo Naranjo Piñera, Sylvia Irene Schmelkes del Valle** y **Patricia Gabriela Vázquez del Mercado Herrera.**

El Director General de Asuntos Jurídicos, **Agustín E. Carrillo Suárez**.- Rúbrica.

(R.- 480012)