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 2015-2016.

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 2015-2016.

Con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 14, 22, 26, 27 fracción VII, 29, 38 fracción VI, 47, 49 de la Ley del Instituto Nacional para la Evaluación de la Educación; Lineamientos para llevar a cabo la evaluación para el ingreso al Servicio Profesional Docente en Educación básica y Educación media superior[1] para el ciclo escolar 2015-2016. LINEE-01-2015, la Junta de Gobierno emite 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 2015-2016.

El presente documento tiene como finalidad establecer los referentes y procedimientos 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) Proceso para la calificación de los sustentantes; 4) Resultado del proceso de evaluación y 5) Integración de las listas de prelación. Se presenta un Anexo técnico con información detallada de algunos de los aspectos técnicos que se consideran en los distintos apartados el documento.

Definición de términos

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

I. **Alto impacto:** Se indica cuando los resultados del instrumento tienen consecuencias importantes para las personas o las instituciones; por ejemplo, en 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 éste se aplica en distintas ocasiones.

IV. **Constructo:** Elaboración teórica formulada para explicar un proceso social, psicológico o educativo y cuya adecuada descripción permite que sea susceptible de ser observable o medible.

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. **Criterios de desempate:** Regla con la cual se determina el orden que ocupan los sustentantes en las listas de prelación, con base en los resultados en los distintos instrumentos que constituyen el proceso de evaluación.

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

VIII. **Desempeño:** Resultado obtenido por el sustentante en un instrumento de evaluación educativa.

IX. **Dificultad de un reactivo:** Indica la proporción de personas que responden correctamente el reactivo de un examen. Entre mayor sea este índice, menor será su dificultad y a mayor dificultad del reactivo, menor será su índice.

---

[1] Se emplean las siglas EB y EMS para referirse a la Educación básica y Educación media superior, respectivamente y SPD para el Servicio Profesional Docente.

X. **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.

XI. **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.

XII. **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.

XIII. **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.

XIV. **Equiparación:** Proceso estadístico que se utiliza para ajustar las puntuaciones de las formas de un mismo instrumento, permite que las puntuaciones de una forma a otra sean utilizadas de manera intercambiable. La equiparación ajusta, por dificultad, las distintas formas que fueron construidas con contenidos y dificultad similar.

XV. **Error estándar de medida:** Desviación estándar de una distribución hipotética de errores de medida de una población.

XVI. **Escala:** Procedimiento para asignar números, puntuaciones o medidas a objetos o sucesos con propiedades específicas a partir de reglas definidas.

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

XVIII. **Especificaciones de tareas evaluativas o reactivos:** Descripción detallada de las características relevantes que se espera tengan los sujetos al sustentar el instrumento de evaluación y que es posible observar a través de las tareas evaluativas o los reactivos. Tienen el papel de guiar a los comités académicos en la elaboración y validación de las tareas evaluativas o los reactivos y que éstos cuenten con los elementos necesarios para construirlos alineados al objeto de medida o constructo que se desea evaluar a través del instrumento.

XIX. **Estándar:** Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación y que son acordados por expertos en evaluación.

XX. **Evaluación:** Acción de emitir juicios de valor que resultan de comparar los resultados de una medición u observación con un referente previamente establecido.

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

XXII. **Indicador:** Instrumento utilizado para determinar, por medio de unidades de medida, el grado de cumplimiento de una característica, cualidad, conocimiento, capacidad, objetivo o meta, empleado para valorar factores que se desean medir.

XXIII. **Instrumento de evaluación:** Técnicas de medición y recolección de datos que suelen 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 pre-establecido.

XXV. **Jueceo:** Método en el cual se utiliza la opinión de expertos (denominados jueces) para determinar, entre otras cosas, la pertinencia de la validez de las tareas evaluativas o los reactivos respecto a un dominio; el establecimiento de estándares o puntos de corte; así como la calificación de reactivos de respuesta construida.

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, 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 las de la población.

XXIX. **Nivel de desempeño:** Criterio conceptual que delimita el marco interpretativo de las puntuaciones obtenidas en un instrumento de evaluación, y que refiere a lo que la persona evaluada es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.

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

XXXI. **Parámetro:** Valor de referencia que permite medir avances y resultados alcanzados en el cumplimiento de objetivos, metas y demás características del ejercicio de una función o actividad.

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 aspirante 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 sujeto cuyo desempeño es evaluado a través de una rúbrica.

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

XXXVI. **Puntuación:** Número de aciertos obtenidos en un instrumento de evaluación.

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

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

XXXIX. **Sesgo:** Error en la medición de un atributo (por ejemplo, conocimiento o habilidad), debido a una variable no controlada, como las diferencias culturales o lingüísticas de las personas evaluadas.

XL. **Sustentante:** Aspirante a ingresar al Servicio Profesional Docente que da respuesta los instrumentos de evaluación que se consideran en el concurso de oposición.

XLI. **Tareas evaluativas:** Unidad básica de medida de un instrumento de evaluación que consiste en la ejecución de una actividad que es susceptible de ser observada.

XLII. **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 cuenta 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 el cual 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 técnico del instrumento.

Con base en los resultados de estos procesos deben identificarse las tareas evaluativas o los reactivos que contribuyen a la calidad métrica del instrumento, los cuales deben incorporarse para la calificación de las personas evaluadas, a fin de estimar con mayor precisión su desempeño.

Para llevar a cabo el análisis de los instrumentos de medición utilizados en los concursos, es necesario que los distintos grupos de sustentantes de las entidades federativas queden equitativamente representados, dado que la cantidad de aspirantes por tipo de evaluación en cada entidad federativa es notoriamente diferente. Para ello, se definirá una muestra de aspirantes 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 a la calificación de los sustentantes. Para conformar dicha muestra, cada entidad federativa contribuirá con 500 aspirantes como máximo, y deberán ser elegidos aleatoriamente. Si hay menos de 500 aspirantes, todos se incluirán en la muestra. Si no se realizara este procedimiento, las decisiones sobre los instrumentos de evaluación, así como en la identificación de los puntos de corte y los estándares de desempeño, se verían fuertemente influenciadas, indebidamente, por el desempeño mostrado por los aspirantes de aquellas entidades que se caracterizan por tener más 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 de opción múltiple:

- Deberán estar organizados jerárquicamente en tres niveles de desagregación (por ejemplo, áreas, subáreas y temas); el primero deberá contar con al menos dos conjuntos de contenidos específicos por evaluar y, cada uno de ellos, deberá tener al menos 20 reactivos efectivos para calificar.
- El segundo nivel de desagregación deberá considerar al menos dos subconjuntos de aspectos a evaluar, y cada uno de ellos deberá tener al menos 10 reactivos efectivos para calificar.
- En el tercer nivel de desagregación, cada aspecto a evaluar deberá contemplar al menos dos contenidos específicos, los cuales deberán estar definidos en términos de especificaciones de reactivos. Cada especificación deberá ser evaluada al menos por un reactivo.
- Las especificaciones de reactivos deberán integrarse por una definición operacional del contenido específico a evaluar, un reactivo ejemplo y la bibliografía en la se sustenta el reactivo.
- Los instrumentos de evaluación de carácter nacional deberán tener, al menos, 80 reactivos efectivos para calificación.
- Los instrumentos complementarios que atienden necesidades específicas de las entidades estatales, deberán tener una longitud igual o mayor a 60 reactivos efectivos para calificar.
- Deberá documentarse el procedimiento que se siguió para determinar la estructura del instrumento y la cantidad de reactivos que conforman el instrumento, a fin de justificar la relevancia (ponderación) de los contenidos específicos evaluados en el mismo.

Exámenes de respuesta construida:

- Deberán estar organizados jerárquicamente en, al menos, dos niveles de desagregación (áreas y subáreas; si fuera el caso, temas); el primero deberá contar, con al menos dos conjuntos de contenidos específicos a evaluar.
- A partir del segundo nivel (o tercer nivel, si fuera el caso) de desagregación, se deberán elaborar las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional.
- A partir de las definiciones operacionales se diseñarán los niveles o categorías de dominio.
- Los distintos niveles o categorías de dominio que se consignen, deberán ser claramente distinguibles entre sí.

Criterios y parámetros estadísticos

Debido a las implicaciones que tienen los resultados de los instrumentos empleados en los concursos de ingreso al Servicio Profesional Docente en EB y EMS, deberán atenderse los siguientes criterios y parámetros estadísticos:

En el caso de los instrumentos de evaluación con 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 0.20.
- Los distractores deberán tener correlaciones punto biserial negativas.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.90.

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

- La confiabilidad del instrumento deberá ser igual o mayor que 0.80.
- La correlación entre cada aspecto evaluado con la puntuación global deberá ser igual o mayor que 0.30.
- El porcentaje de acuerdos inter-jueces deberá ser mayor o igual a 70%.
- El porcentaje de acuerdos intra-jueces deberá ser mayor o igual a 80% considerando al menos 5 medidas repetidas seleccionadas al azar.

Si en algún instrumento de evaluación no se llegara a cumplir con estos parámetros estadísticos y la falta de reactivos comprometiera la estructura diseñada del instrumento de evaluación que fue aprobada por el Consejo Técnico, podrán considerarse los siguientes parámetros estadísticos:

En el caso de los instrumentos de evaluación con 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.

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

- La correlación 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 0.70.
- El porcentaje de acuerdos inter-jueces deberá ser mayor o igual a 60%.
- El porcentaje de acuerdos intra-jueces deberá ser mayor o igual a 70% considerando al menos 5 medidas repetidas seleccionadas al azar.

> *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 la estructura del instrumento que fue aprobada por el Consejo Técnico.*

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 utilizarán para evaluar a los sustentantes de los concursos de oposición de ingreso al SPD, 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 de 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 docente adecuado. 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 etapas, las cuales se describen a continuación:

Primera etapa

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, para orientar el trabajo de los comités académicos en el desarrollo de los descriptores específicos de cada instrumento. Para todos los instrumentos se utilizarán tres niveles de desempeño posibles: Nivel I (N I), Nivel II (N II) y Nivel III (N III). 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 (N I)	Dominio insuficiente de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño docente.
Nivel II (N II)	Dominio suficiente y organizado de los conocimientos y habilidades, contemplados en el instrumento, que se juzgan indispensables para un adecuado desempeño docente.
Nivel III (N III)	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 utilizarlas en una diversidad de situaciones didácticas.*

Segunda etapa

En esta etapa 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 y poder identificar cuál es la puntuación requerida para que el sustentante alcance un determinado nivel o estándar de desempeño.

Los insumos que tendrán los Comités académicos como referentes para el desarrollo de esta actividad, será la documentación que describe la estructura de los instrumentos, sus especificaciones y los ejemplos de reactivos incluidos en las mismas. En todos los casos, el primer punto de corte se establecerá a partir de lo que los expertos definan como la ejecución típica o esperable de un sustentante hipotético, mínimamente aceptable, para cada nivel de desempeño (NII o NIII). Para ello los expertos reunidos en los Comités académicos, deberán determinar, para cada pregunta considerada, cuál es la probabilidad de que dichos sustentantes hipotéticos las respondan 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, los Comités académicos, considerando el conjunto de reactivos que, en cada caso el sustentante hipotético es capaz de responder, deberán describir los conocimientos y habilidades específicos que están implicados en cada nivel de desempeño, en términos de lo que éste conoce y es capaz de hacer.

Tercera etapa

En la tercera etapa 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, una vez que se cuente con los resultados alcanzados por los sustentantes, 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 alcanzara cada nivel de desempeño (II y III) previamente definido, 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).

La tercera etapa se llevará 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 aspirantes, los puntos de corte serán los definidos en la segunda etapa.

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 la determinación de los puntos de corte correspondientes.

3. Proceso para la calificación de los sustentantes

Todos los sustentantes que participen en el Concurso de Oposición al Ingreso al SPD 2015-2016, en EB y EMS 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 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 programa de evaluación implica la aplicación 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 de 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.
- 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

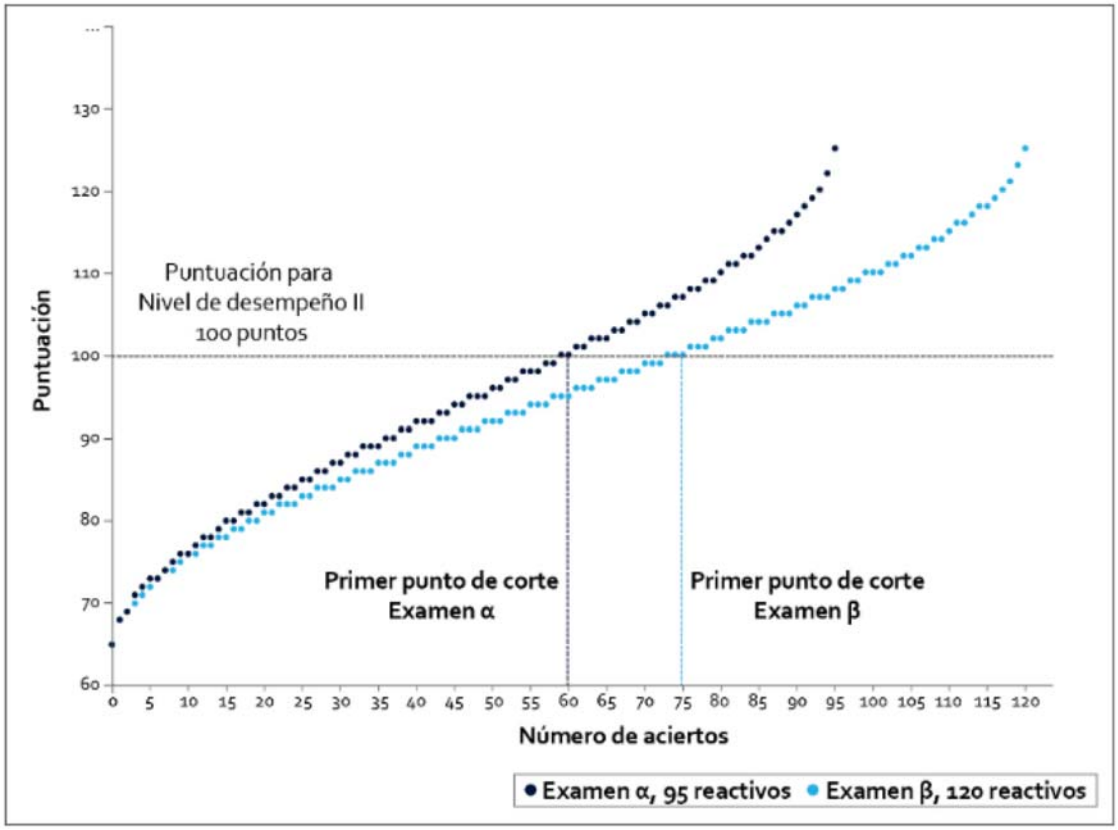
En cada plan de evaluación es indispensable definir el tipo de escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. 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.

Por esta razón, dado que los instrumentos de evaluación utilizados en el proceso de evaluación de los concursos son de carácter criterial, toma especial relevancia emplear una escala de calificación diferente en las que tradicionalmente se reportan los resultados educativos en México como la escala de 5 a 10 (Santiago, 2014), de 0 a 10 bien de 0% a 100%, con la finalidad de evitar interpretaciones equívocas de los resultados, por ejemplo, si se obtiene una puntuación del 50% de aciertos y con esta puntuación se ubica en el nivel II de desempeño, podría afirmarse erróneamente que “se está aprobando, aun cuando se reprobó en la prueba”.

El escalamiento que se llevará a cabo, permitirá construir una métrica común para todos los instrumentos de evaluación 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 o más en esta escala (cuyo rango va de va de 60 a 170 puntos[2]), representa, en todos los instrumentos, que se ha alcanzado un nivel de desempeño II, al menos. Es decir, que se cuenta con un “dominio suficiente y organizado de los conocimientos y habilidades (contemplados en el instrumento) que se juzgan indispensables para un adecuado desempeño docente”.

[2] Pueden encontrarse ligeras variaciones en este rango debido a que la escala es aplicable a múltiples instrumentos con características muy diversas, tales como las longitudes, tipos de instrumentos y su nivel de precisión, diferencias entre los puntos de corte que atienden a las particularidades de los contenidos que se evalúan, entre otras. Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

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 diferentes. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



4. Resultado del proceso de evaluación

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 el resultado del proceso de la evaluación que permite establecer la idoneidad de los sustentantes, deberán integrarse los resultados de todos los instrumentos de evaluación sustentados, bajo el criterio de que:

> *El sustentante Idóneo será aquel que obtenga, al menos, el nivel de desempeño II (N II)* ***en todos y cada uno*** *de los instrumentos de evaluación que constituyen el proceso de evaluación, según se define en los lineamientos del concurso*

Cada sustentante conocerá el resultado integrado de todo el proceso de evaluación, así como los resultados de 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 la 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 el proceso de evaluación. Los grupos de desempeño son el primer criterio de ordenamiento para la integración de las listas de prelación.

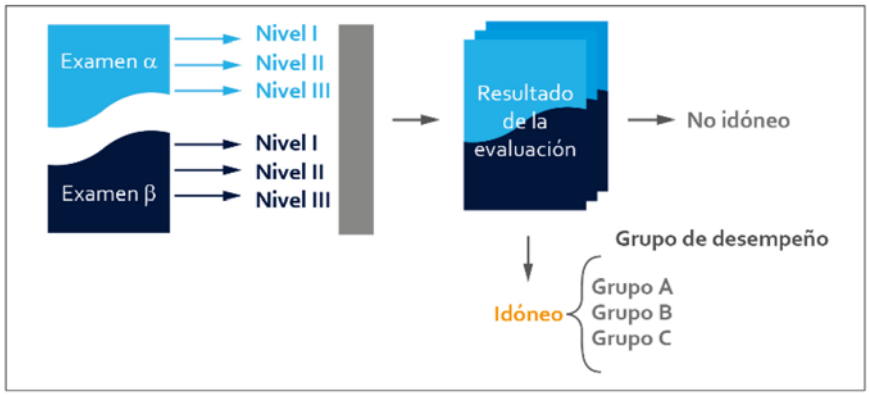
Como ejemplo, los grupos de desempeño en los procesos de evaluación que constan de dos únicos instrumentos, se conformarán de la manera siguiente:

El **primer grupo de desempeño (A)**, se conformará con aquellos sustentantes que alcancen el nivel de desempeño III (N III) 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 (N II) en uno de los exámenes y el nivel de desempeño III (N III) en el otro examen.

El **tercer grupo de desempeño (C),** se conformará por los aspirantes que alcancen el nivel de desempeño N II en los dos de los exámenes.

En la siguiente figura se representan los grupos de desempeño derivados del ejemplo:



Para los casos en los que el proceso de evaluación considere tres o más instrumentos, los grupos se deberán integrar con la misma lógica de ordenamiento en función del desempeño observado en cada uno de ellos. A continuación se presentan las tablas 2a, 2b y 2c 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	
Grupos de desempeño	**Descripción**
A	En los dos exámenes obtuvo el nivel de desempeño III
B	En un examen obtuvo el nivel de desempeño III y en el otro el nivel de desempeño II
C	En los dos exámenes obtuvo en nivel de desempeño II

Educación básica

Presentan dos exámenes los aspirantes a plazas docentes y técnico docentes: 1) Examen disciplinar, correspondiente a la convocatoria por la cual concursa y 2) el instrumento común denominado Examen de conocimientos y Habilidades para la práctica docente.

Tabla 2b. Grupos de desempeño con tres exámenes

Criterios para formar parte de un grupo de desempeño	
Grupos de desempeño	**Descripción**
A	En los tres exámenes obtuvo el nivel de desempeño III
B	En dos exámenes obtuvo el nivel de desempeño III y en el otro nivel de desempeño II
C	En un examen obtuvo el nivel de desempeño III y en los otros dos el nivel de desempeño II
D	En los tres exámenes obtuvo el nivel de desempeño II

Educación media superior

Presentan cuatro exámenes los aspirantes a plazas docentes del componente disciplinar. Los instrumentos son: 1) Conocimientos disciplinares, 2) Habilidades docentes, 3) Plan de clase y 4) Exprese.

Tabla 2c. Grupos de desempeño con cuatro exámenes

Criterios para formar parte de un grupo de desempeño	
Grupos de desempeño	**Descripción**
A	En los cuatro exámenes obtuvo el nivel de desempeño III
B	En tres exámenes obtuvo el nivel de desempeño III y en el otro nivel de desempeño II
C	En dos exámenes obtuvo el nivel de desempeño III y en los otros dos el nivel de desempeño II
D	En un examen obtuvo el nivel de desempeño III y en los otros tres el nivel de desempeño II
E	En los cuatro exámenes obtuvo el nivel de desempeño II

Educación básica

Presentan tres exámenes los aspirantes a plazas docentes con requerimientos particulares de acuerdo a la especificación de la Entidad federativa correspondiente: 1) Examen disciplinar 2) Examen de conocimientos y Habilidades para la práctica docente y 3) Examen complementario, de acuerdo a lo que se indique en la convocatoria

Educación media superior

Presentaran tres exámenes los aspirantes a plazas docentes y técnico docentes de las disciplinas asociadas al componente profesional técnico. Los instrumentos son: 1) Habilidades docentes, 2) Plan de clase y 3) Exprese.

5. Integración de las listas de prelación

Las listas de prelación se integrarán sólo con sustentantes que alcancen 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, después C, 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.

De manera excepcional, si después de llevar a cabo el proceso anterior para generar las listas de prelación se observan empates entre 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 EB y EMS, referidos en los párrafos anteriores.

Criterios específicos de ordenamiento para EB

Jerarquía de los instrumentos de evaluación para los docentes

En EB los grupos de desempeño se definen de acuerdo con lo estipulado en la Tabla 2a, cuando el proceso de evaluación considera sólo dos exámenes:

1º Examen de conocimientos y habilidades para la práctica docente

2º Examen 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 EB

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Preescolar*		Primaria		Secundaria			Especial	Educación física	Inglés
		General	Indígena	General	Indígena	General	Técnica	Telesecundaria**			
Examen de Conocimientos y Habilidades para la Práctica Docente	Intervención didáctica	1	1	2	2	2	2	1	2	2	2
	Aspectos curriculares	2	2	1	1	1	1	2	1	1	1
Examen 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, la prioridad 1 es la Intervención didáctica y la 2 es Aspectos curriculares.

**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, 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 curriculares.

Jerarquía de los instrumentos de evaluación para técnicos docentes

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

1º Examen de conocimientos y habilidades para la práctica docente

2º Examen 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 EB

Contenidos específicos de primer nivel	Contenidos de segundo nivel	Maestro de taller de Lectura y Escritura. Preescolar, Primaria y Secundaria	Acompañante de Música. Preescolar	Maestra de enseñanza Artística. Primaria	Maestro de Taller. Primaria	Maestro de Música. Indígena	Maestro de Taller. Indígena	Acompañante de Música. Educación Especial.	Maestro Taller. Educación Especial.	Maestro de aula de medios. Secundaria
Examen de Conocimientos y Habilidades para la Práctica Docente	Intervención didáctica	2	2	2	2	2	2	2	2	2
	Aspectos curriculares	1	1	1	1	1	1	1	1	1
Examen de Habilidades Intelectuales y Responsabilidades Ético Profesionales	Compromiso ético	3	3	3	3	3	3	3	3	3
	Mejora profesional	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

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 indican en la Tabla 2b, ya que se consideran tres exámenes:

1º Examen de conocimientos y habilidades para la práctica docente

2º Examen de habilidades intelectuales y responsabilidades ético profesionales

3° 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

Los grupos de desempeño para evaluar a los docentes se definen de acuerdo con lo estipulado en la Tabla 2c, cuando el proceso de evaluación considera cuatro exámenes:

1º Examen de conocimientos sobre contenidos disciplinares

2º Examen de conocimientos sobre habilidades docentes

3º Plan de clase

4º Exprese

La jerarquización de los contenidos específicos de segundo nivel de los instrumentos de evaluación, se 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 conocimientos sobre contenidos disciplinares y el Examen de conocimientos sobre habilidades docentes.

Los instrumentos de evaluación de Plan de clase y Exprese no ingresan al tercer nivel de ordenamiento debido a que se evalúan a través de una rúbrica.

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

Los grupos de desempeño se indican en la Tabla 2b, ya que consideran 3 exámenes:

1º Examen de conocimientos sobre habilidades docentes

2º Plan de clase

3º Exprese

La jerarquización de los contenidos específicos de segundo nivel de los instrumentos de evaluación, se 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 conocimientos sobre habilidades docentes.

Los instrumentos de evaluación de Plan de clase y Exprese no ingresan al tercer nivel de ordenamiento debido a que se evalúan a través de una rúbrica.

Anexo técnico

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. La suma de las probabilidades se expresará en una puntuación esperada del examen para cada juez. 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 aspirantes una vez que un instrumento de evaluación es administrado.

Procedimiento

En el cuerpo del documento se señalaron tres fases para el establecimiento de puntos de corte de los niveles de desempeño. 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 las fases anteriores. Las dos preguntas son:

a) ¿Cuál es el mínimo nivel de conocimientos o habilidades que un aspirante 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 aspirantes 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 de los instrumentos.

Adicionalmente, se debe contar con la distribución de los aspirantes 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).

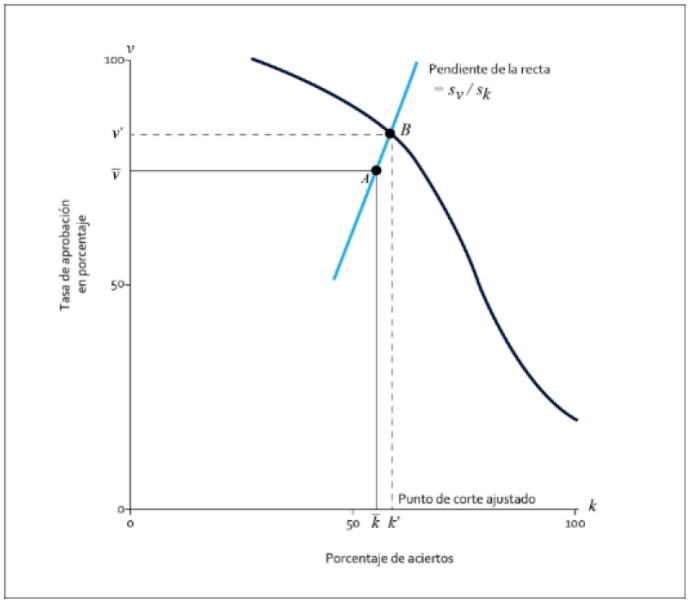
2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

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 se llevará a cabo a partir de las puntuaciones crudas (cantidad de aciertos) de los sustentantes, y se obtendrá una métrica común para todos los instrumentos de evaluación, que va de 60 a 170 puntos aproximadamente, ubicando el primer punto de corte (nivel de desempeño II) para todos 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. **De no implementarla, para cada instrumento se tendría que estimar el error estándar de medida para todas y cada una de las puntuaciones de la escala** (estándares 2.3 y 2.4 de los Estándares para las Pruebas Educativas y Psicológicas de la American Educational Research Association et. al., 2014)**.**

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.

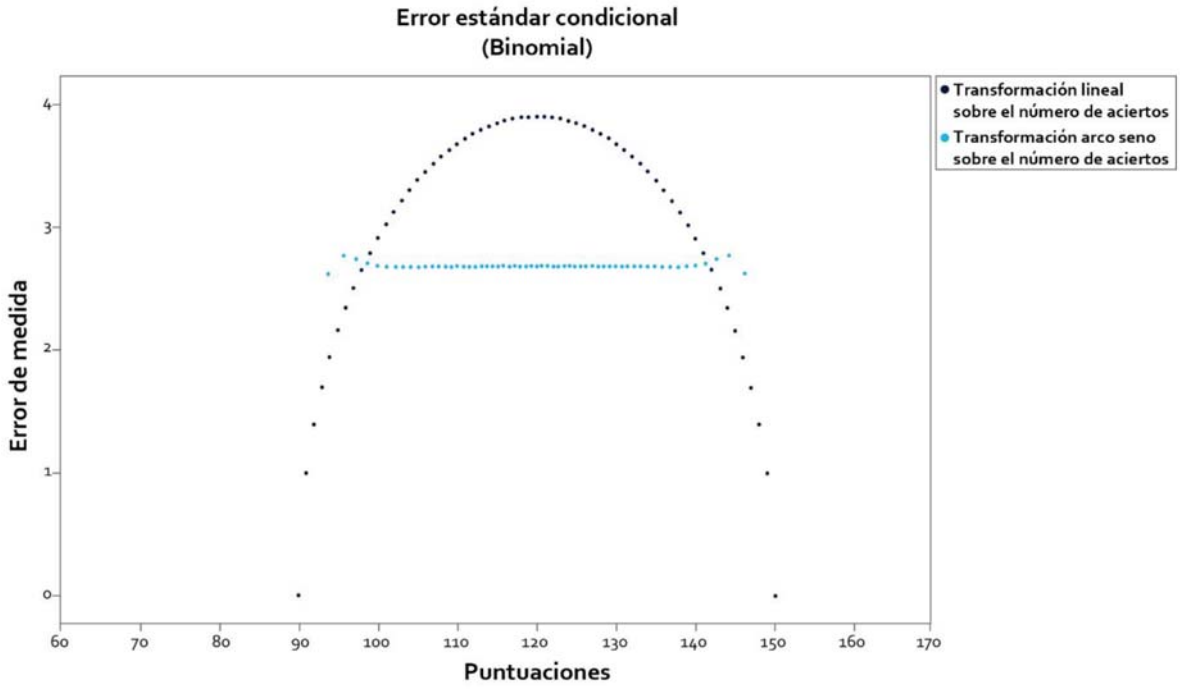
Además, el escalamiento permitirá reportar, en el informe técnico del instrumento, **un solo valor para el error estándar de medida en la escala para cada instrumento de evaluación** y con ello atender a los estándares internacionales anteriormente citados.

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 en este anexo.

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 aspirante en el instrumento de evaluación. Los reactivos se calificarán como correctos o incorrectos de acuerdo con la clave de respuesta correspondiente. Si un aspirante 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 aspirantes considerando que *K* sea la máxima puntuación que se pueda obtener en el instrumento de evaluación.

Como se observa en la gráfica (Won-Chan, Brennan y Kolen, 2000), con excepción de los valores extremos, el error estándar de medición se estabiliza a lo largo de la distribución de las puntuaciones observadas, a diferencia de la transformación lineal de las puntuaciones crudas.



Para estabilizar la varianza de los errores estándar de medición a lo largo de la escala, se utilizará la función *c:*

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

Donde:

i se refiere a un aspirante

k_i es el número de respuestas correctas que el aspirante *i* obtuvo en el examen

K es el número de reactivos del examen

Procedimiento para la transformación lineal

La puntuación mínima aceptable que los aspirantes deben tener para ubicarse en el nivel de desempeño II (N II) en los instrumentos de evaluación, se ubicará en el valor 100. Para determinarla se empleará la siguiente ecuación:

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

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 aspirante para ubicarlo en el nivel de desempeño II.

El valor de Q tomará los valores 60 o de 80 dependiendo de la confiabilidad del instrumento. Para confiabilidades igual o mayores a 0.90, Q tomará el valor 80 y, si es menor a 0.90 tomará el valor 60. Lo anterior implica que los extremos de la escala puedan 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 aspirante (*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_{Ai}). Las puntuaciones de los contenidos específicos (P_{Ai}) 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 (P_{A1i}) que tenga la mayor relevancia, 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:

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

El resultado se redondeará al entero inmediato anterior con el criterio de que puntuaciones con cinco décimas 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 una 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 de la *j-ésima* contenido específico será:

$$P_{Aji} = P_i - \sum_j P_{Aji} \quad (5)$$

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 defino el valor de un cociente en donde el denominador tome el valor de cero. En este caso, el puntaje deberá registrase 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* (Kendal y Suart 2000), que calcula los errores estándar de medición condicionales. Para incluir la confiablidad del instrumento de medición se usa un modelo de error binomial, Keats propone que el cálculo del error estándar condicional de medición sea:

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

Donde X es una variable aleatoria asociada a los puntajes, α es el coeficiente de confiabilidad de Cronbach (KR-20) y KR21 es el coeficiente de Kuder-Richardson.

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.

Para los puntajes que se les aplique la equiparación, $x_e = b1x + b0$, con *b1* como pendiente y *b0* como ordenada al origen; el procedimiento es análogo, 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:

$$\sigma(x_e) = b1 * \sigma(x)$$

Procedimiento para la Equiparación de puntuaciones

La equiparación de las formas de un instrumento deberá realizarse utilizando el método de equiparación lineal de Levine (Kolen y Brennan, 2010), 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, $w_1 + w_2 = 1$ y $w_1, w_2 \geq 0$. Para este proceso se utilizará

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

y

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

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:

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

Donde s denota la población sintética y

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

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

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

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

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

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

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 γ´s se expresan de la siguiente manera:

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

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

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.

Finalmente, como ya se indicó en el cuerpo del documento, se deberán considerar 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*).

Referencias

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCM). (2014). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd 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) p. 147-152.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2: 130–44.

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

Kolen, M. & Brennan, R. (2010). *Test Equating, Scaling, and Linking*. New York, NY: Springer Verlag.

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

Santiago, P. et. al. (2014). *Revisiones de la OCDE sobre la Evaluación en Educación.* Instituto Nacional para la Evaluación de la Educación (INEE).

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

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.

México, D.F., a seis de abril de dos mil quince.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Quinta Sesión Extraordinaria de dos mil quince, celebrada el seis de abril de dos mil quince. Acuerdo número **SEJG/5-15/03,R**. La Consejera Presidenta, **Sylvia Irene Schmelkes del Valle**.- Rúbrica.- Los Consejeros: **Eduardo Backhoff Escudero**, **Teresa Bracho González**, **Gilberto Ramón Guevara Niebla**, **Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 410293)