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 emisión de resultados para llevar a cabo la evaluación del desempeño del personal con funciones de dirección y supervisión en Educación Básica en el ciclo escolar 2017-2018.

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 EMISIÓN DE RESULTADOS PARA LLEVAR A CABO LA EVALUACIÓN DEL DESEMPEÑO DEL PERSONAL CON FUNCIONES DE DIRECCIÓN Y SUPERVISIÓN EN EDUCACIÓN BÁSICA EN EL CICLO ESCOLAR 2017-2018.

El presente documento está dirigido a las autoridades educativas que en el marco de sus atribuciones implementan evaluaciones que, por la naturaleza de sus resultados, regula el Instituto Nacional para la Evaluación de la Educación (INEE), en especial las referidas al Servicio Profesional Docente (SPD) que son desarrolladas por la Coordinación Nacional del Servicio Profesional Docente (CNSPD).

Con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 7, fracción X de la Ley General del Servicio Profesional Docente; 22, 28, fracción X, 38, fracciones VI, IX y XXII de la Ley del Instituto Nacional para la Evaluación de la Educación; en los Lineamientos para llevar a cabo la evaluación del desempeño del personal con funciones de Dirección y Supervisión en Educación Básica en el ciclo escolar 2017-2018 (LINEE-05-2017), la Junta de Gobierno aprueba los siguientes criterios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados para llevar a cabo la evaluación del desempeño del personal con funciones de dirección y supervisión en Educación Básica en el ciclo escolar 2017-2018.

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. Su contenido se organiza de la siguiente manera:

Primera sección: Sobre la evaluación del desempeño para el ciclo escolar 2017-2018

Incorpora cinco apartados: 1) Características generales de los instrumentos para evaluar el desempeño del personal con funciones de dirección y supervisión; 2) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 3) Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3; 4) Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global.

Segunda sección: Sobre la evaluación del desempeño de quienes será su segunda oportunidad.

En la parte final se presenta un Anexo técnico con información detallada de algunos de los aspectos técnicos que se consideran en el documento.

Definición de términos

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

I. **Alto impacto:** Se indica cuando los resultados de un 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.

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. **Cuestionario:** Tipo de instrumento de evaluación que sirve para recolectar información sobre actitudes, conductas, opiniones, contextos demográficos o socioculturales, entre otros.

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

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

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. **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 una misma persona en un mismo instrumento que tienden a distribuirse alrededor de un puntaje verdadero. El puntaje verdadero siempre es desconocido porque ninguna medida puede ser una representación perfecta de un puntaje verdadero.

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. Constituye el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XIX. **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.

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

XXI. **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.

XXII. **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.

XXIII. **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.

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

XXV. **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.

XXVI. **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.

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

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

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

XXX. **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.

XXXI. **Porcentaje de acuerdos intra-jueces:** Medida del grado en que el mismo juez, a través de dos o más mediciones repetidas a los mismos sujetos que evalúa, coincide en la puntuación asignada al desempeño de los sujetos, evaluados a través de una rúbrica.

XXXII. **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.

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

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

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

XXXVI. **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.

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

XXXVIII. **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.

Primera sección.

Evaluación del desempeño del personal con funciones de dirección y supervisión en Educación Básica, 2017-2018

1. Características generales de los instrumentos para evaluar el desempeño del personal con funciones de dirección y supervisión

La evaluación del desempeño es un proceso integrado que incluye varios instrumentos que dan cuenta de los diferentes aspectos que se describen en los Perfiles, parámetros e indicadores establecidos por la autoridad educativa. A continuación, se describen sucintamente cada uno de los instrumentos considerados en cada etapa del proceso.

Personal con funciones de dirección

(Director, Subdirector académico de secundaria, Coordinador de actividades académicas de secundaria y Subdirector de gestión de secundaria)

Etapa 1. Informe de responsabilidades profesionales

Esta etapa está constituida por dos instrumentos de evaluación, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales del personal con funciones de dirección, de sus procesos de aprendizaje y mejora permanente en el ejercicio de su función, así como de su colaboración en el trabajo de la escuela y de la zona escolar:

a) Cuestionario de autoevaluación, respondido por el personal con funciones de dirección

b) Cuestionario para su autoridad inmediata, quien proporcionará la información relativa al nivel de cumplimiento de las responsabilidades profesionales del personal con funciones de dirección

Etapa 2. Proyecto de gestión escolar del personal con funciones de dirección

El proyecto de gestión escolar del personal con funciones de dirección es un instrumento que permite evaluar el desempeño de su gestión directiva a través de una muestra genuina de su práctica profesional. Consiste en la elaboración de un plan de trabajo, el desarrollo de la estrategia del plan y la selección de evidencias para elaborar un texto de análisis y reflexión sobre su gestión directiva. Está constituido por tres momentos:

Momento 1. Elaboración de un plan de trabajo de gestión

Momento 2. Desarrollo del plan de trabajo de gestión

Momento 3. Análisis y reflexión de su gestión directiva

Etapa 3. Examen de conocimientos curriculares y de normatividad para el personal con funciones de dirección

Este instrumento evalúa los conocimientos, capacidades y habilidades de este personal para afrontar y resolver diversas situaciones de la práctica profesional y propiciar la mejora de las prácticas de los docentes y del funcionamiento de la escuela. Los principales aspectos a evaluar son los conocimientos de la función, las acciones de intervención, la reflexión de su práctica, la aplicación de la normatividad y el trabajo colaborativo.

Personal con funciones de supervisión

(Supervisor y Jefe de sector)

Etapa 1. Informe de responsabilidades profesionales

Esta etapa está constituida por dos instrumentos de evaluación, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales del personal con funciones de supervisión, de sus procesos de aprendizaje y mejora permanente en el ejercicio de su función, así como de su colaboración en el trabajo de las escuelas de su zona o sector escolar:

a) Cuestionario de autoevaluación, respondido por el personal con funciones de supervisión

b) Cuestionario para su autoridad inmediata, quien proporcionará la información relativa al nivel de cumplimiento de las responsabilidades profesionales del personal con funciones de supervisión

Etapa 2. Proyecto de asesoría y acompañamiento del personal con funciones de supervisión

El proyecto de asesoría y acompañamiento del personal con funciones de supervisión es un instrumento que permite evaluar su desempeño a través de una muestra genuina de su práctica profesional. Consiste en la elaboración de un plan de asesoría y acompañamiento para la atención de una prioridad educativa, el desarrollo del plan y la selección de evidencias, así como la elaboración de un texto de análisis en donde reflexione sobre su intervención. Está constituido por tres momentos:

Momento 1. Elaboración de un plan de trabajo

Momento 2. Desarrollo del plan de trabajo

Momento 3. Análisis y reflexión de su práctica profesional

Etapa 3. Examen de conocimientos curriculares y de normatividad para el personal con funciones de supervisión

Este instrumento evalúa los conocimientos, capacidades y habilidades de este personal para afrontar y resolver diversas situaciones de la práctica profesional y propiciar la mejora de las prácticas del conjunto de escuelas adscritas a su zona o sector escolar. Los principales aspectos a evaluar son los conocimientos de la función, las acciones de intervención, la reflexión de su práctica, la aplicación de la normatividad y el trabajo colaborativo.

2. 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 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 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 sustentantes con el nivel de precisión requerido para los propósitos de la evaluación, los instrumentos 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 estar organizados jerárquicamente en tres niveles de desagregación: áreas, subáreas y temas, en donde:
  - Cada instrumento debe contar con al menos dos áreas.
  - 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).

Cuestionarios que constituyen la etapa 1:

- En una matriz se deben identificar los indicadores y variables de interés, así como definir sus componentes.
- El contenido debe estar organizado jerárquicamente en dos niveles de desagregación, en donde el primero debe contar, como mínimo, con dos conjuntos de contenidos específicos.

Criterios y parámetros estadísticos

Los instrumentos empleados para la evaluación del desempeño 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:

I. 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.

II. 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 la 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. Estas mediciones deberán aportarse antes de emitir la calificación definitiva de los sustentantes, a fin de salvaguardar la confiabilidad de la decisión.

III. En el caso de los cuestionarios que constituyen la Etapa *1. Informe de responsabilidades profesionales*, para cada una de las escalas que los constituyen:

- La correlación entre cada reactivo con la puntuación global de la escala deberá ser igual o mayor que 0.20.
- La confiabilidad del constructo medido a través de la escala debe ser igual o mayor que 0.80.

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.

3. Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3

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 la evaluación del desempeño, 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 sustentantes 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 cuatro niveles de desempeño posibles: Nivel I (N I), Nivel II (N II), Nivel III (N III) y Nivel IV (N IV). Los descriptores genéricos para los diferentes grupos de instrumentos y cada nivel se indican en las Tablas 1a y 1b para el caso de quienes ejercen funciones de dirección.

Tabla 1a. Descriptores genéricos de los niveles de desempeño para el instrumento Proyecto de gestión escolar del personal con funciones de dirección

Nivel de desempeño	Descriptor
Nivel I (N I)	El personal con funciones de dirección demuestra carencia de conocimientos para explicar la tarea de la escuela, desconoce la mayoría de los componentes del currículo y de los elementos de trabajo en el aula relacionados con el logro de los aprendizajes de los alumnos; además menciona parcialmente los mecanismos para la mejora de las prácticas docentes. Enlista acciones que podrían implementarse para la construcción de ambientes de trabajo en la escuela que posibiliten la sana convivencia o la inclusión educativa, pero omite la diversidad cultural y lingüística de la comunidad; menciona ambiguamente la Normalidad Mínima de Operación Escolar y las prácticas que conservan la integridad y seguridad de los alumnos en el aula y en la escuela. Señala de forma imprecisa algunos ejemplos de participación de la comunidad en las tareas de la escuela.
Nivel II (N II)	El personal con funciones de dirección demuestra conocimientos básicos para explicar la tarea de la escuela, menciona los componentes del currículo y los elementos de trabajo en el aula relacionados con el logro de los aprendizajes de los alumnos, aunque no abunda en los mecanismos implementados para la mejora de las prácticas docentes. Menciona algunas acciones que se implementan para la construcción de ambientes de trabajo en la escuela que posibiliten la sana convivencia y la inclusión educativa, sin considerar apropiadamente la diversidad cultural y lingüística de la comunidad; adicionalmente menciona la Normalidad Mínima de Operación Escolar y únicamente enlista algunas prácticas que conservan la integridad y seguridad de los alumnos en el aula y en la escuela. Señala la importancia de la participación de la comunidad en la tarea educativa de la escuela.
Nivel III (N III)	El personal con funciones de dirección proporciona evidencias sólidas para explicar la tarea de la escuela, explica los componentes del currículo, los mecanismos implementados para la mejora de las prácticas docentes y los elementos de trabajo en el aula, pero sin relacionarlos con el logro de los aprendizajes de los alumnos. Fundamenta de forma limitada las acciones que se implementan para la construcción de ambientes de trabajo en la escuela que posibiliten la sana convivencia y la inclusión educativa, sin considerar adecuadamente la diversidad cultural y lingüística de la comunidad; también busca cumplir con la Normalidad Mínima de Operación Escolar y reconocer las prácticas que conservan la integridad y seguridad de los alumnos en el aula y en la escuela. Busca la participación de la comunidad y de otras instituciones en la tarea educativa de la escuela.
Nivel IV (N IV)	El personal con funciones de dirección demuestra conocimientos y habilidades sólidos para explicar la tarea fundamental de la escuela, explica ampliamente los componentes del currículo y los elementos de trabajo en el aula en relación con el logro de los aprendizajes de los alumnos, así como los mecanismos implementados para la mejora de las prácticas docentes. Fundamenta las acciones que se implementan para la construcción de ambientes de trabajo en la escuela que posibilitan la sana convivencia y la inclusión educativa, tomando en cuenta la diversidad cultural y lingüística de la comunidad; además, asegura la Normalidad Mínima de Operación Escolar y reconoce lo fundamental de las prácticas que conservan la integridad y seguridad de los alumnos en el aula y en la escuela. Busca de manera prioritaria la participación de la comunidad y de otras instituciones en la tarea educativa de la escuela.

Tabla 1b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos curriculares y de normatividad para el personal con funciones de dirección

Nivel de desempeño	Descriptor
Nivel I (N I)	El personal con funciones de dirección demuestra carencia de conocimientos en la gestión de tareas de organización para la mejora de la calidad educativa. Desconoce la mayoría de los componentes del currículo y su relación con los elementos del aprendizaje de los alumnos. Identifica alguna práctica educativa para atender niñas, niños y adolescentes con Necesidades Educativas Especiales, en alguna situación de vulnerabilidad, en riesgo de reprobación, rezago o deserción escolar, identifica acciones que garantizan su seguridad en la escuela. Desconoce aspectos que promueven ambientes para favorecer el aprendizaje, la sana convivencia o la inclusión educativa y reconoce acciones poco apropiadas para fomentar la vinculación de la diversidad cultural y lingüística con la tarea educativa de la escuela. Omite acciones que fomenten la colaboración de las familias y de la comunidad. Desconoce estrategias adecuadas para la búsqueda, selección y uso de información de distintas fuentes para el apoyo de su desarrollo profesional, además reconoce los principios filosóficos, los fundamentos legales o las finalidades de la educación pública mexicana, sin vincularlos con su función directiva.
Nivel II (N II)	El personal con funciones de dirección demuestra conocimientos básicos en la gestión de tareas de organización para la mejora de la calidad educativa, conoce algunos de los componentes del currículo pero los relaciona parcialmente con los elementos del aprendizaje de los alumnos, identifica pocas prácticas educativas para atender niñas, niños y adolescentes con necesidades educativas especiales, en alguna situación de vulnerabilidad, en riesgo de reprobación, rezago o deserción escolar, y sólo reconoce algunas acciones para garantizar su integridad o seguridad en la escuela. Reconoce pocos aspectos que promueven ambientes para favorecer el aprendizaje, la sana convivencia o la inclusión educativa y reconoce acciones que fomentan escasamente la vinculación de la diversidad cultural y lingüística con la tarea educativa de la escuela. Identifica escasas acciones que fomentan la colaboración de las familias, la comunidad y otras instituciones. Emplea estrategias poco adecuadas para la búsqueda, selección y uso de información de distintas fuentes para el apoyo de su desarrollo profesional, aunque no parte de referentes teóricos, además reconoce en general los principios filosóficos, los fundamentos legales o las finalidades de la educación pública mexicana.
Nivel III (N III)	El personal con funciones de dirección demuestra conocimientos sólidos en la gestión de tareas de organización para la mejora de la calidad educativa, conoce la mayoría de los componentes del currículo y su relación con el aprendizaje de los alumnos, identifica estrategias generales para orientar a los docentes en su intervención y conoce algunas prácticas educativas para atender niñas, niños y adolescentes con necesidades educativas especiales, en alguna situación de vulnerabilidad, en riesgo de reprobación, rezago o deserción escolar y para garantizar parcialmente su integridad y seguridad en la escuela. Reconoce algunos elementos que promueven ambientes para favorecer el aprendizaje, la sana convivencia o la inclusión educativa y reconoce acciones que sirven para vincular la diversidad cultural y lingüística con la tarea educativa de la escuela. Identifica algunas acciones que promueven la colaboración de las familias, la comunidad y otras instituciones para el funcionamiento de la zona escolar y en el trabajo con otros directivos. Emplea estrategias para la búsqueda, selección y uso de información de distintas fuentes disponibles en su contexto para el apoyo de su desarrollo profesional partiendo de referentes teóricos, además reconoce los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana dentro de su función directiva.

Nivel IV (N IV)	El personal con funciones de dirección demuestra conocimientos y habilidades sólidos en la gestión de tareas de organización para la mejora de la calidad educativa, conoce los componentes del currículo y su relación con el aprendizaje de los alumnos, selecciona estrategias eficaces para orientar a los docentes en su intervención e identifica plenamente las prácticas educativas para atender niñas, niños y adolescentes con necesidades educativas especiales, en alguna situación de vulnerabilidad, en riesgo de reprobación, rezago o deserción escolar y para garantizar su integridad y seguridad en los diferentes espacios de la escuela. Detecta los ambientes que favorecen el aprendizaje, la sana convivencia y la inclusión educativa, vinculando la diversidad cultural y lingüística con la tarea educativa de la escuela. Identifica acciones pertinentes para fomentar la colaboración de las familias, la comunidad y otras instituciones aportando estrategias al funcionamiento eficaz de la zona escolar y en el trabajo con otros directivos. Elige estrategias adecuadas para la búsqueda, selección y uso de información de distintas fuentes, así como materiales impresos y Tecnologías de la Información y la Comunicación disponibles en su contexto para el apoyo de su desarrollo profesional partiendo de referentes teóricos confiables y pertinentes, además reconoce los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana que dan sustento al ejercicio de su función directiva.

Para el caso de quienes ejercen funciones de supervisión, los descriptores genéricos se indican en las Tablas 2a y 2b.

Tabla 2a. Descriptores genéricos de los niveles de desempeño para el instrumento Proyecto de asesoría y acompañamiento del personal con funciones de supervisión

Nivel de desempeño	Descriptor
Nivel I (N I)	El personal con funciones de supervisión proporciona evidencias donde demuestra que carece de habilidades para asumir la función de la supervisión escolar para la mejora de la calidad educativa. Posee escasas habilidades para organizar el trabajo de las escuelas y responsabiliza a cada escuela para el cumplimiento de la Normalidad Mínima de Operación Escolar. Hace un resumen ambiguo sobre su práctica profesional empleando algún medio de consulta para enriquecer su desarrollo en diferentes ámbitos. Observa que los ambientes sean favorables para el aprendizaje, la sana convivencia o la inclusión, del mismo modo se limita a un reporte con las acciones implementadas para el cuidado de la seguridad de los alumnos en las escuelas. Identifica algún caso de diversidad cultural y lingüística de los alumnos implicados en algunos procesos educativos, pero no establece comunicación estrecha con las familias y las comunidades.
Nivel II (N II)	El personal con funciones de supervisión proporciona evidencias donde demuestra que asume sólo parcialmente la función de la supervisión escolar para la mejora de la calidad educativa, pues la vincula con pocos propósitos, enfoques o contenidos educativos, y únicamente enlista prácticas educativas que propician aprendizajes. Posee algunas habilidades para organizar el trabajo de las escuelas y busca que cada escuela cumpla con la Normalidad Mínima de Operación Escolar. Propone algunas pautas para que cada escuela desarrolle el sistema de asesoría y acompañamiento, estrategias para la mejora de la gestión y ocasionalmente busca comunicación entre las escuelas y autoridades educativas, así como de otras instituciones. Verifica que los ambientes sean favorables para el aprendizaje, la sana convivencia o la inclusión, del mismo modo coteja que se efectúen algunas acciones implementadas para el cuidado de la seguridad de los alumnos en las escuelas. Identifica casos específicos de diversidad cultural y lingüística de los alumnos implicados en algunos procesos educativos propiciando comunicación con las familias y las comunidades.
Nivel III (N III)	El personal con funciones de supervisión proporciona algunas evidencias donde asume la función de la supervisión escolar para la mejora de la calidad educativa, las vincula con algunos propósitos, enfoques y contenidos educativos y analiza de manera independiente las prácticas educativas que propician aprendizajes. Posee algunas habilidades para organizar el trabajo de las escuelas y busca cumplir con la Normalidad Mínima de Operación Escolar. Propone el sistema de asesoría y acompañamiento, estrategias para la mejora de la gestión y busca comunicación entre las escuelas y autoridades educativas, así como de otras instituciones. Verifica que los ambientes sean favorables para el aprendizaje, la sana convivencia y la inclusión, del mismo modo coteja las acciones implementadas para el cuidado de la integridad y la seguridad de los alumnos en las escuelas. Reconoce la diversidad cultural y lingüística de los alumnos implicados en algunos procesos educativos propiciando comunicación con familias, comunidades y otras instituciones en la tarea educativa de las escuelas de la zona escolar.

Nivel IV (N IV)	El personal con funciones de supervisión proporciona evidencias sólidas donde asume la función de la supervisión escolar para la mejora de la calidad educativa, la vincula coherentemente con los propósitos, enfoques y contenidos educativos y analiza profundamente las prácticas educativas que propician aprendizajes. Posee diversas habilidades para organizar el trabajo de las escuelas, contribuye al cumplimiento de la Normalidad Mínima de Operación Escolar. Organiza el sistema de asesoría y acompañamiento, las estrategias para la mejora de la gestión y establece vínculos entre las escuelas y autoridades educativas, así como de otras instituciones que apoyen la tarea educativa. Reflexiona sistemáticamente sobre su práctica profesional como medio para mejorarla utilizando diferentes medios para enriquecer su desarrollo profesional. Gestiona ambientes favorables para el aprendizaje, la sana convivencia y la inclusión, del mismo modo establece estrategias para el cuidado de la integridad y la seguridad de los alumnos en las escuelas. Considera la diversidad cultural y lingüística de los alumnos y su vinculación con los procesos educativos estableciendo estrategias de colaboración con las familias, las comunidades y otras instituciones en la tarea educativa de las escuelas de la zona escolar.

Tabla 2b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos curriculares y de normatividad para el personal con funciones de supervisión

Nivel de desempeño	Descriptor
Nivel I (N I)	El personal con funciones de supervisión demuestra carecer de conocimientos sobre las funciones de la supervisión escolar para la mejora de la calidad educativa. Conoce parcialmente estrategias para organizar y gestionar el trabajo entre las escuelas, así como procesos para dar seguimiento al sistema de asesoría y acompañamiento para las escuelas. Identifica referentes teóricos que no dan fundamento para tomar decisiones que permitan mejorar su práctica profesional. Reconoce la existencia de los principios filosóficos, los fundamentos legales o las finalidades de la educación pública mexicana, pero de forma desvinculada con el ejercicio de su función y sin reflejar la gestión de ambientes para el aprendizaje, la sana convivencia o la inclusión educativa en las escuelas. Identifica alguna estrategia para el seguimiento parcial a casos de violencia, abuso o maltrato infantil. Reconoce algún ejemplo de diversidad cultural o lingüística de los alumnos, aunque estén desvinculados con sus procesos educativos y conoce de forma limitada acciones que favorezcan la colaboración de las familias o las comunidades en la tarea educativa de las escuelas de la zona escolar.
Nivel II (N II)	El personal con funciones de supervisión demuestra conocimientos básicos en el reconocimiento de las funciones de la supervisión escolar para la mejora de la calidad educativa, identifica algunos propósitos, enfoques y contenidos educativos independientemente del análisis de las prácticas educativas. Conoce pocas estrategias para organizar y gestionar el trabajo entre las escuelas para implementar la calidad educativa, reconoce algún proceso para dar seguimiento al sistema de asesoría y acompañamiento para las escuelas, aunque no siempre considera los lineamientos establecidos. Reconoce los principios filosóficos, los fundamentos legales o las finalidades de la educación pública mexicana de forma desvinculada con el ejercicio de su función y la gestión de ambientes para el aprendizaje, la sana convivencia o la inclusión educativa en las escuelas de la zona, además identifica alguna estrategia para el seguimiento a casos de violencia, abuso o maltrato infantil en colaboración con los directivos escolares. Reconoce ejemplos de diversidad cultural o lingüística de los alumnos, aunque estén desvinculados con sus procesos educativos.

Nivel III (N III)	El personal con funciones de supervisión demuestra conocimientos sólidos en las funciones de la supervisión escolar para la mejora de la calidad educativa, identifica los propósitos, enfoques y contenidos educativos independientemente de las prácticas educativas que propician aprendizajes. Conoce algunas estrategias para organizar y gestionar el trabajo entre las escuelas para mejorar la calidad educativa, reconoce los procesos para dar seguimiento al sistema de asesoría y acompañamiento para las escuelas. Emplea algunos referentes teóricos que dan fundamento para tomar decisiones que permitan mejorar su práctica profesional. Reconoce los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana en el ejercicio de su función que favorecen la gestión de ambientes para el aprendizaje, la sana convivencia o la inclusión educativa en las escuelas, además reconoce estrategias para el seguimiento a casos de violencia, abuso o maltrato infantil en colaboración con los directivos escolares. Identifica la diversidad cultural y lingüística de los alumnos y su vinculación con procesos educativos donde colaboran las familias y las comunidades en la tarea educativa de las escuelas.
Nivel IV (N IV)	El personal con funciones de supervisión demuestra conocimientos y habilidades sólidos en las funciones de la supervisión escolar para la mejora de la calidad educativa, las relaciona con propósitos, enfoques y contenidos educativos y reconoce las prácticas educativas que propician aprendizajes. Conoce diversas estrategias para organizar y gestionar óptimamente el trabajo entre las escuelas para mejorar la calidad educativa, reconoce los procesos para dar seguimiento al sistema de asesoría y acompañamiento para las escuelas. Identifica diversos referentes teóricos que dan fundamento para tomar decisiones que permitan mejorar su práctica profesional, tomados de diversas fuentes. Domina los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana en el ejercicio de su función que propician la adecuada gestión de ambientes favorables de aprendizaje, la sana convivencia y la inclusión educativa en las escuelas de la zona, además reconoce la importancia de establecer estrategias para el seguimiento a casos de violencia, abuso o maltrato infantil en colaboración con los directivos escolares. Reconoce la diversidad cultural y lingüística de los alumnos y su vinculación con procesos educativos donde colaboran las familias, las comunidades y otras instituciones en la tarea educativa de las escuelas.

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 Instituto determinará el procedimiento a seguir para el establecimiento de los puntos de corte correspondientes, atendiendo al marco jurídico aplicable.

4. Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global

A continuación, se presentan dos subapartados, en el primero se describen los procedimientos para calificar los resultados de los sustentantes en cada instrumento[1] en cada etapa; mientras que en el segundo se detallan los procedimientos para la obtención del resultado global.

4.1 Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen las etapas del proceso de evaluación

4.1.1 Con relación a los instrumentos considerados en las etapas 2 y 3

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

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 sustentante 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 sustentantes (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).

[1] En el caso en que el sustentante **no presente alguno** de los instrumentos de evaluación de las etapas 2 y 3 o el cuestionario de autoevaluación de la etapa 1, 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. Para el caso en que el sustentante no presente ninguno de los instrumentos de evaluación de las etapas 2 y 3 ni el cuestionario de autoevaluación de la etapa 1, su resultado global será "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: no presentó", asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos que constituyen el proceso de evaluación del desempeño. En el caso en que la autoridad inmediata no responda el cuestionario que le corresponde de la etapa 1, el resultado en ese instrumento será "SI: sin información".

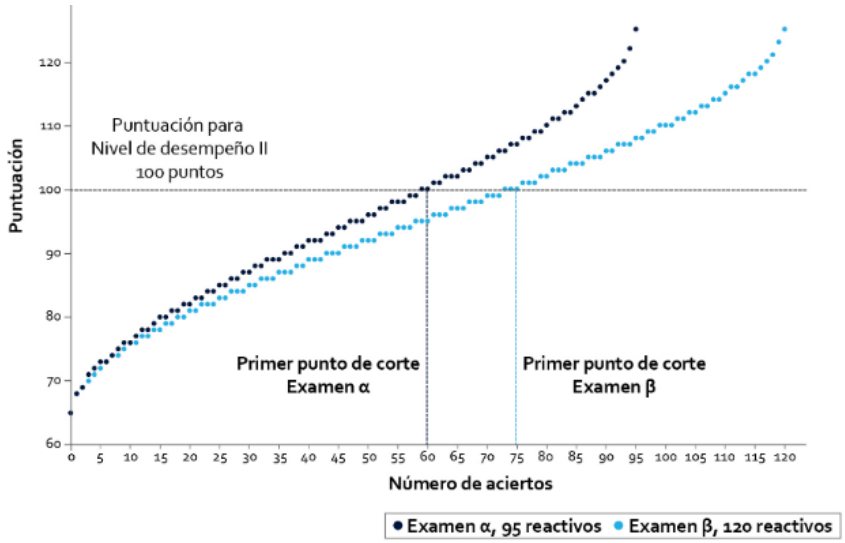
Escala utilizada para reportar los resultados

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 los instrumentos de las etapas 2 y 3 de este proceso de evaluación, permitirá construir una métrica común. 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 los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos[2]).

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 instrumento el punto de corte refiera a número de aciertos diferente. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



4.1.2 Con relación a los cuestionarios que integran la Etapa *1. Informe de responsabilidades profesionales*

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos, se integrará la información y se definirán cuatro categorías que indicarán el nivel de cumplimiento del sustentante en las responsabilidades profesionales de su función[3]. Cada una de estas categorías tendrá asociada una cantidad de puntos que, como posteriormente se indicará, se adicionará a la puntuación total ponderada, considerando el siguiente orden:

[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, los 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; por otra parte, para realizar el escalamiento, el sustentante debe, al menos, haber alcanzado un acierto en el examen; en caso contrario, se reportará como cero y obtendrá N I. Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

[3] Para mayores detalles sobre el procedimiento para el escalamiento de las puntuaciones de los cuestionarios, la integración de la información y la asignación de niveles de cumplimiento en la etapa 1, consultar el anexo técnico.

NI: 0 puntos

NII: 1 punto

NIII: 2 puntos

NIV: 3 puntos

Cada cuestionario contribuirá con el 50% de la puntuación de la etapa 1, de tal forma que, en caso de faltar las respuestas de alguno de los dos cuestionarios, la puntuación de la etapa será igual a la puntuación que aporta el cuestionario del que se cuente con información.

En ningún caso, por sí mismo, la omisión de alguno de los dos cuestionarios que considera esta etapa de la evaluación **será causal de un resultado Insuficiente**. Lo anterior porque se trata de reconocer y estimular la participación genuina de los sustentantes y autoridades superiores.

4.2 Resultado global y procedimiento para la conformación de los grupos de desempeño

4.2.1 El resultado global

Para determinar el resultado global de la calificación de los sustentantes, deberán integrarse los resultados de los instrumentos considerados en las tres etapas que conforman el diseño de la evaluación, conforme a los siguientes criterios:

1) Sustentar los dos instrumentos que constituyen las etapas 2 y 3

2) Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3

Cuando no se cumpla con los criterios 1 y 2, no aplicarán los numerales 3, 4 y 5

3) Una vez que se verifica el cumplimiento de los criterios 1 y 2, se calcula la puntuación total ponderada del sustentante, es decir, se pondera[4] el resultado obtenido en los dos instrumentos de las etapas 2 y 3 bajo el siguiente esquema:

   a. Etapa 2. Proyecto de gestión escolar del personal con funciones de dirección, 60% *(el nombre del proyecto varía para el personal con funciones de supervisión)*

   b. Etapa 3. Examen de conocimientos curriculares y de normatividad para el personal con funciones de dirección, 40% (el nombre del examen varía para el personal con funciones de supervisión)

4) Se adiciona el resultado obtenido en la etapa 1, de acuerdo con el nivel de cumplimiento alcanzado: NI (0 puntos), NII (1 punto), NIII (2 puntos), o bien NIV (3 puntos).

5) Se asigna el resultado global de la evaluación, que integra los resultados parciales de todo el proceso.

4.2.2 La conformación de los grupos de desempeño

El resultado "Suficiente"

Para alcanzar al menos un resultado suficiente en la evaluación, se deben cumplir los siguientes criterios:

- Sustentar los dos instrumentos que constituyen las etapas 2 y 3
- Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3
- Obtener al menos 100 puntos en la escala de calificación global

Los **grupos de desempeño** estarán conformados únicamente por los sustentantes que obtengan, al menos, un resultado "Suficiente" en la evaluación:

Criterios para formar parte de un grupo de desempeño	
Grupo de desempeño	Puntuación en escala de calificación global
Suficiente	Al menos 100[5] puntos
Bueno	Al menos PC_2G puntos
Destacado	Al menos PC_3G puntos

[4] Se traduce como la cantidad de puntos en escala INEE multiplicado por 0.60 y 0.40, respectivamente. Para mayores detalles sobre el algoritmo para el cálculo de la puntuación global, consultar el anexo técnico.

[5] PC_1G siempre es igual a 100, toda vez que el primer punto de corte en los instrumentos considerados en las etapas 2 y 3 siempre es 100. Para mayores detalles sobre el algoritmo para el cálculo de los puntos de corte en la escala de calificación global, consultar el anexo técnico.

El resultado "Insuficiente"

En los siguientes casos se asignará el resultado "Insuficiente" y, por lo tanto, el sustentante **no formará parte de los grupos de desempeño**, **pero recibirá la retroalimentación que corresponda**:

- No sustente los dos instrumentos que constituyen las etapas 2 y 3.
- **No obtenga** al menos NII en por lo menos uno de los dos instrumentos que constituyen las etapas 2 y 3.
- No obtenga **al menos** 100 puntos en la escala de calificación global.

En los dos primeros casos no se dará puntuación global al sustentante.

En los tres casos los sustentantes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan presentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

El resultado "No se presentó a la evaluación"

Para el caso en que el sustentante no presente ninguno de los instrumentos de las etapas 2 y 3 considerados en el diseño de la evaluación, ni el cuestionario de autoevaluación de la etapa 1, en el resultado de la evaluación se indicará: "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: No presentó". Asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos, aun cuando su autoridad inmediata haya respondido el cuestionario que le corresponde de la etapa 1.

Sobre los resultados de la evaluación

El resultado de la evaluación, tanto para los resultados "Insuficientes", como de aquellos que forman parte de un grupo de desempeño ("Suficiente", "Bueno" o "Destacado"), aportará información relevante para diseñar programas y acciones de capacitación, formación y acompañamiento.

Segunda sección.

Evaluación del desempeño en su segunda oportunidad del personal con funciones de dirección y supervisión en Educación Básica

De conformidad con la Ley General del Servicio Profesional Docente, esta evaluación del desempeño en su segunda oportunidad es obligatoria y deberá llevarse a cabo en un plazo no mayor de doce meses después de haberse presentado la primera evaluación.

Serán sujetos a una segunda oportunidad de evaluación del desempeño exclusivamente quienes ejercen funciones de dirección y supervisión que obtuvieron resultado insuficiente en su primera evaluación del desempeño.

La calificación global se estimará siguiendo el mismo modelo de calificación desarrollado en los presentes criterios técnicos (véase la primera sección). Se considerarán los resultados obtenidos en su primera oportunidad con base en las siguientes equivalencias:

Equivalencias para la etapa 2

Se recuperará la información de los resultados que el sustentante haya obtenido en su primera oportunidad en los siguientes instrumentos de evaluación, con base en sus funciones:

Personal con funciones de dirección	Personal con funciones de supervisión
● *Ruta de mejora argumentada* ● *Expediente de evidencias de la función de dirección*	● *Plan de trabajo argumentado* ● *Expediente de evidencias de la función de supervisión*

Las reglas de equivalencias serán las siguientes:

Resultado obtenido en Ruta de mejora argumentada o Plan de trabajo argumentado	***Resultado obtenido en Expediente de evidencias de la función de dirección o Expediente de evidencias de la función de supervisión***	**Resultado asignado para la etapa 2**
NII, NIII o NIV	*NII, NIII o NIV*	El nivel de desempeño más alto que haya alcanzado en cualquiera de los dos instrumentos
En cualquier resultado cuya combinación de los dos instrumentos sea: NP o NI con NP, NI, NII, NIII o NIV		Debe presentar el instrumento que le corresponda de la etapa

Equivalencias para la etapa 3

Se recuperará la información de los resultados que el sustentante haya obtenido en su primera oportunidad en los siguientes instrumentos de evaluación, con base en sus funciones:

Personal con funciones de dirección	**Personal con funciones de supervisión**
• *Examen de conocimientos y habilidades directivas*	• *Examen de conocimientos y habilidades de las funciones de supervisión*

Las reglas de equivalencias serán las siguientes:

Resultado obtenido en ***Examen de conocimientos y habilidades directivas o*** ***Examen de conocimientos y habilidades de las funciones de supervisión***	**Resultado asignado para la etapa 3**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NI o NP	Debe presentar el instrumento que le corresponda de la etapa

Finalmente, cualquier situación no prevista en los presentes criterios técnicos será analizada por la Junta de Gobierno 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* y toda vez que los cuestionarios que constituyen la etapa 1 de este proceso tienen como finalidad recabar información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función profesional, y **únicamente** pueden ser considerados para **adicionar puntos al sustentante en su calificación global, la cual está en función de los resultados alcanzados en los instrumentos que constituyen las etapas 2 y 3**, es fundamental señalar que, en ningún caso, **se puede considerar solamente un instrumento** para integrar la calificación de los sustentantes conforme al diseño de la evaluación, es decir:

Ninguna decisión que tenga consecuencias importantes sobre los individuos o instituciones, se basará únicamente en los resultados de sólo un instrumento de evaluación, por lo cual, deberán considerarse otras fuentes confiables de información que incrementen la validez de las decisiones que se tomen.

Lo anterior debido a que la evidencia empírica que resulte del análisis psicométrico de los instrumentos de la segunda y tercera etapa de la evaluación del desempeño del personal con funciones de dirección y supervisión debe mostrar que, una vez que éstos fueron aplicados, cumplen con los criterios técnicos establecidos por el Instituto, de esta forma 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 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 ni el centro de trabajo 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 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 calificación que se asigna 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 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 del sustentante en cada aspecto a evaluar para emitir su calificación definitiva 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 del punto 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 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:

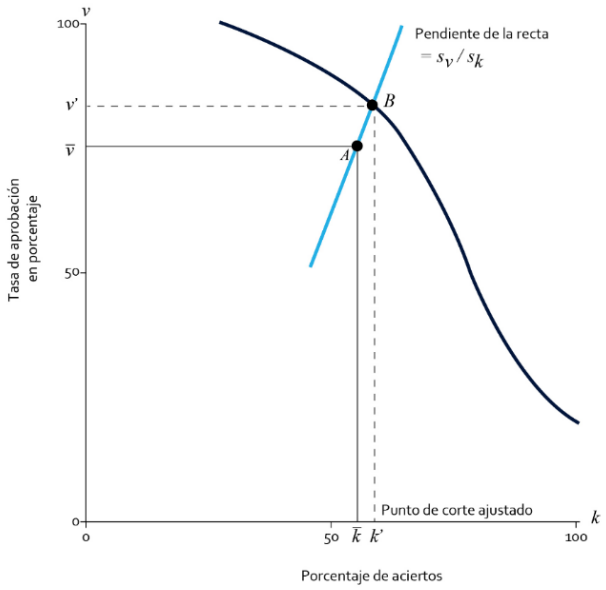
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas ($\bar{k}$, $\bar{v}$), (ver siguiente figura).

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: $\boldsymbol{v = \left(s_v/s_k\right)\left(k - \bar{k}\right) + \bar{v}}$.

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 de los instrumentos considerados en las etapas 2 y 3

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 II) 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).

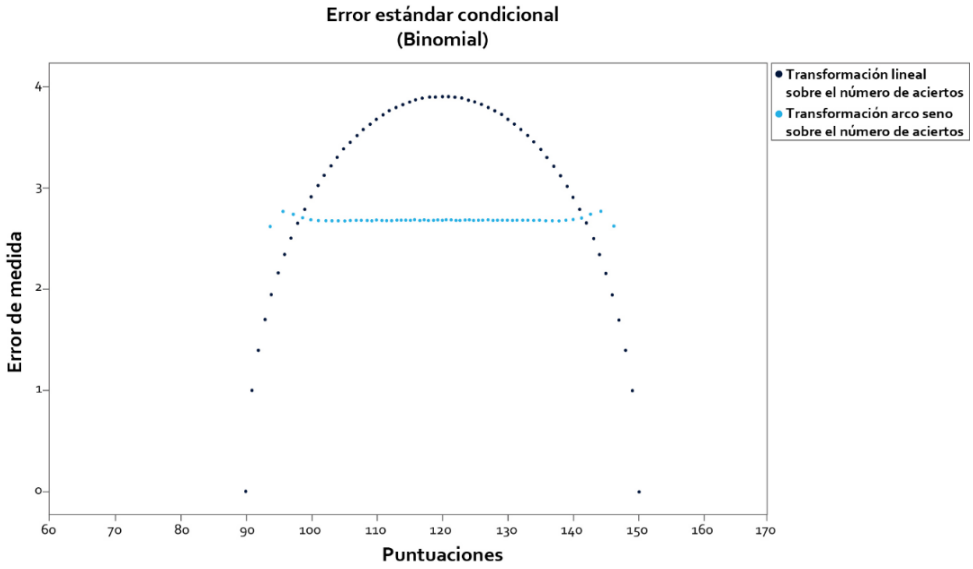
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

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 N I.

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

Como se comentó, 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 II (N II) en los instrumentos de evaluación, se ubicará en el valor 100. Para determinarla se empleará la siguiente ecuación:

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

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 II.

El valor de Q dependerá 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 (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 puntuaciones 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 1 (P_{Aji}), 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 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, 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 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, el punto 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.

$\boldsymbol{\alpha}$ 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:

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

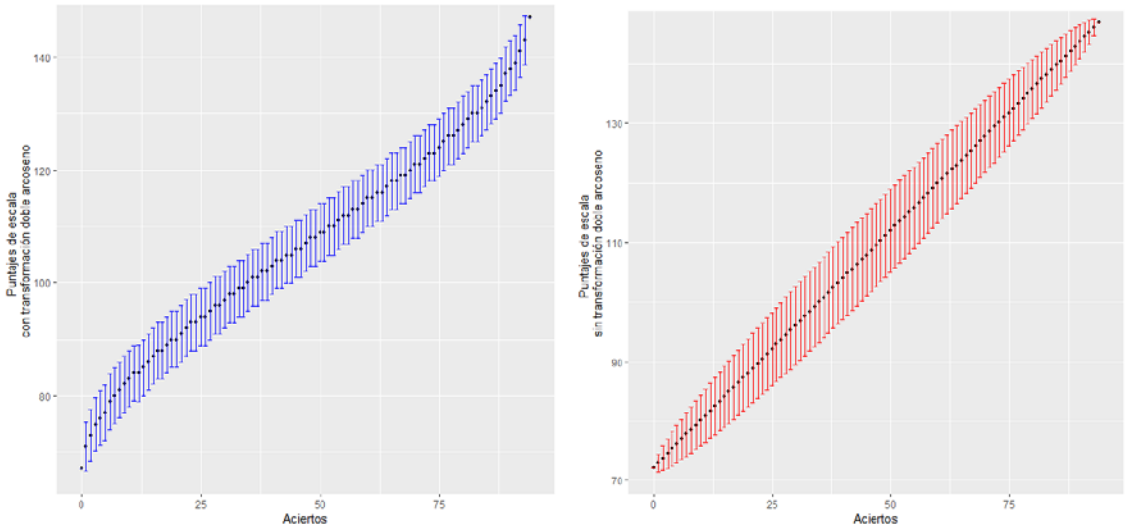
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

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.

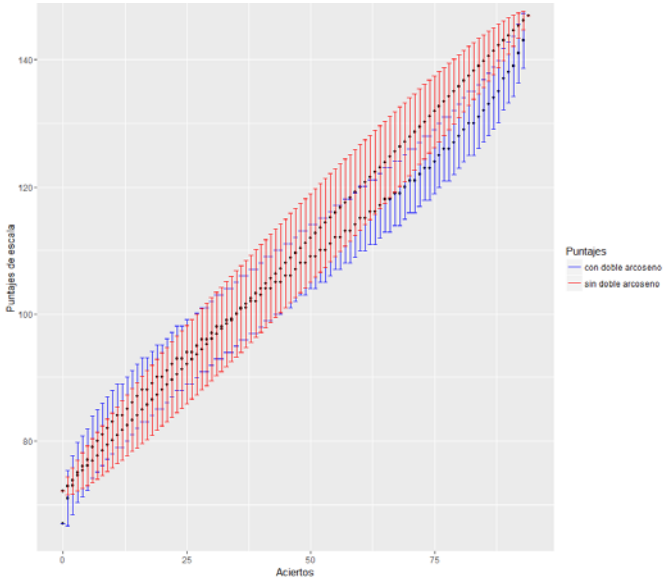
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 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 II (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

Como ya se indicó en el cuerpo del documento, 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.

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

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.

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).

Procedimiento para el escalamiento de las puntuaciones de los cuestionarios de la etapa 1

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos.

La escala de puntuaciones de cada cuestionario se ubicará en el intervalo [0, 50], si un cuestionario no es presentado se le asignará una puntuación de cero. Ambos cuestionarios serán escalados utilizando el modelo de crédito parcial. Para que el rango de puntuaciones vaya de 0 a 50, las puntuaciones que se obtengan con el modelo se escalarán linealmente y se redondearán al entero más próximo, utilizando el criterio de que puntuaciones con cinco décimas o más, suben al siguiente entero.

De esta forma, la puntuación alcanzada en la etapa 1 será calculada como la suma de las puntuaciones de ambos cuestionarios, por lo que se ubicará en el intervalo [0, 100].

La asignación del nivel de cumplimiento en la etapa 1 y la cantidad de puntos que se adicionan a la puntuación total del sustentante, será con base en la siguiente tabla:

Suma de las puntuaciones de ambos cuestionarios	Nivel de cumplimiento	Puntos que se adicionan
De 0 a 25	NI	0
De 26 a 50	NII	1
De 51 a 75	NIII	2
De 76 a 100	NIV	3

Algoritmo para el cálculo de la puntuación global

Una vez que se ha verificado que el sustentante presentó los dos instrumentos que constituyen las etapas 2 y 3 del proceso de evaluación y que obtuvo al menos NII en por lo menos uno de ellos, se procede a calcular la puntuación global con base en el siguiente esquema:

Etapa 2. Proyecto de gestión escolar del personal con funciones de dirección, 60% *(el nombre del proyecto varía para el personal con funciones de supervisión)*

Etapa 3. Examen de conocimientos curriculares y de normatividad para el personal con funciones de dirección, 40% *(el nombre del examen varía para el personal con funciones de supervisión)*

$$G_i = 0.60 * P_{1i} + 0.40 * P_{2i} + P_{Ei}$$

$G_i = Puntuación\ global\ que\ alcanza\ el\ sustentante\ i\ en\ la\ evaluación$

$P_{1i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Proyecto\ de\ gestión\ escolar\ del\ personal\ con\ funciones\ de\ dirección\ (el\ nombre\ del\ proyecto\ varía\ para\ el\ personal\ con\ funciones\ de\ supervisión)$

$P_{2i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ de\ conocimientos\ curriculares\ y\ de\ normatividad\ para\ el\ personal\ con\ funciones\ de\ dirección\ (el\ nombre\ del\ examen\ varía\ para\ el\ personal\ con\ funciones\ de\ supervisión)$

$P_{Ei} = 0,1,2,3\ (Puntuación\ que\ se\ adiciona\ con\ base\ en\ el\ resultado\ del\ sustentante\ i\ en\ la\ etapa\ 1)$

Algoritmo para el cálculo de los puntos de corte en la escala de calificación global

Los puntos de corte en la escala global se calcularán considerando los puntos de corte establecidos en los instrumentos utilizados en las etapas 2 y 3, con base en el siguiente algoritmo:

$$PC_iG = 0.60 * PC_iP + 0.40 * PC_iE$$

$i = 1,2,3$

$PC_iG = Punto\ de\ corte\ i\ en\ la\ escala\ de\ calificación\ global$

$PC_iP = Punto\ de\ corte\ i\ establecido\ en\ el\ Proyecto\ de\ gestión\ escolar\ del\ personal\ con\ funciones\ de\ dirección\ (el\ nombre\ del\ proyecto\ varía\ para\ el\ personal\ con\ funciones\ de\ supervisión)$

$PC_iE = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ de\ conocimientos\ curriculares\ y\ de\ normatividad\ para\ el\ personal\ con\ funciones\ de\ dirección\ (el\ nombre\ del\ examen\ varía\ para\ el\ personal\ con\ funciones\ de\ supervisión)$

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.

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.

Bentler, P. M. (2006). EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

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.

Brennan, R. L. (2012). Scaling PARCC Assessments: Some considerations and a synthetic data example en: http://parcconline.org/about/leadership/12-technical-advisory-committee

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

Downing, SM (2004). Reliability: On the reproducibility of assessment data. Med Educ; 38(9):1006-1012. 21

Holland, P. W., & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), Statistical models for test equating, scaling, and linking (pp. 89–107). New York, NY: Springer

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. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Masters, Geoff (1982). A Rasch model for Partial Credit Scoring. Psychometrika-vol. 47, No. 2.

Muñiz, José (2003): Teoría clásica de los test. Ediciones pirámide, Madrid.

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

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

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

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

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

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

Shun-Wen Chang (2006) Methods in Scaling the Basic Competence Test, *Educational and Psychological* Measurement, 66 (6) 907-927

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

Stemler, E. & Tsai, J. (2008). *Best Practices in Interrater Reliability Three Common Approaches in* Best practices in quantitative methods (pp. 29–49). SAGE Publications, Inc.

Thompson, Bruce ed. (2003): Score reliability. Contemporary thinking on reliability issues. SAGE Publications, Inc.

Wilson, Mark (2005). Constructing measures. An ítem response modeling approach. Lawrence Erlbaum Associates, Publishers.

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.

Wu, Margaret & Adams, Ray (2007). Applying the Rasch Model to Psycho-social measurement. A practical approach. Educational measurement solutions, Melbourne.

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

Ciudad de México, a veintiocho de septiembre de dos mil diecisiete. - Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil diecisiete, celebrada el veintiocho de septiembre de dos mil diecisiete. Acuerdo número **SOJG/09-17/06,R.** El Consejero Presidente, **Eduardo Backhoff Escudero**.- Rúbrica.- Los Consejeros: **Gilberto Ramón Guevara Niebla, Sylvia Irene Schmelkes del Valle, Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 457557)

CRITERIOS técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Básica en el ciclo escolar 2017-2018.

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 EMISIÓN DE RESULTADOS PARA LLEVAR A CABO LA EVALUACIÓN DEL DESEMPEÑO DEL PERSONAL DOCENTE Y TÉCNICO DOCENTE EN EDUCACIÓN BÁSICA EN EL CICLO ESCOLAR 2017-2018.

El presente documento está dirigido a las autoridades educativas que en el marco de sus atribuciones implementan evaluaciones que, por la naturaleza de sus resultados, regula el Instituto Nacional para la Evaluación de la Educación (INEE), en especial las referidas al Servicio Profesional Docente (SPD) que son desarrolladas por la Coordinación Nacional del Servicio Profesional Docente (CNSPD).

Con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 7, fracción X de la Ley General del Servicio Profesional Docente; 22, 28, fracción X, 38, fracciones VI, IX y XXII de la Ley del Instituto Nacional para la Evaluación de la Educación; en los Lineamientos para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Básica y Media Superior en el ciclo escolar 2017-2018 (LINEE-04-2017), la Junta de Gobierno aprueba los siguientes criterios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Básica en el ciclo escolar 2017-2018.

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. Su contenido se organiza de la siguiente manera:

Primera sección: Sobre la evaluación del desempeño para el ciclo escolar 2017-2018.

Incorpora cinco apartados: 1) Características generales de los instrumentos para evaluar el desempeño del personal docente y técnico docente; 2) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 3) Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3; 4) Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global.

Segunda sección: Sobre la evaluación del desempeño de quienes será su segunda o tercera oportunidad.

En la parte final se presenta un Anexo técnico con información detallada de algunos de los aspectos técnicos que se consideran en el documento.

Definición de términos

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

I. **Alto impacto:** Se indica cuando los resultados de un 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.

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. **Cuestionario:** Tipo de instrumento de evaluación que sirve para recolectar información sobre actitudes, conductas, opiniones, contextos demográficos o socioculturales, entre otros.

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

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

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. **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 una misma persona en un mismo instrumento que tienden a distribuirse alrededor de un puntaje verdadero. El puntaje verdadero siempre es desconocido porque ninguna medida puede ser una representación perfecta de un puntaje verdadero.

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. Constituye el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XIX. **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.

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

XXI. **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.

XXII. **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.

XXIII. **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.

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

XXV. **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.

XXVI. **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.

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

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

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

XXX. **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.

XXXI. **Porcentaje de acuerdos intra-jueces:** Medida del grado en que el mismo juez, a través de dos o más mediciones repetidas a los mismos sujetos que evalúa, coincide en la puntuación asignada al desempeño de los sujetos, evaluados a través de una rúbrica.

XXXII. **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.

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

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

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

XXXVI. **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.

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

XXXVIII. **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.

Primera sección.

Evaluación del desempeño del personal docente y técnico docente en Educación Básica, 2017-2018

1. Características generales de los instrumentos para evaluar el desempeño del personal docente y técnico docente

La evaluación del desempeño es un proceso integrado que incluye varios instrumentos que dan cuenta de los diferentes aspectos que se describen en los Perfiles, parámetros e indicadores establecidos por la autoridad educativa. A continuación, se describen sucintamente cada uno de los instrumentos considerados en cada etapa del proceso.

Etapa 1. Informe de responsabilidades profesionales

Esta etapa está constituida por dos instrumentos de evaluación, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función del personal docente y técnico docente, así como identificar las fortalezas y aspectos a mejorar en su práctica:

a) Cuestionario de autoevaluación, respondido por el docente o técnico docente

b) Cuestionario para su autoridad inmediata, quien proporcionará la información relativa al nivel de cumplimiento de las responsabilidades profesionales del docente o técnico docente

Etapa 2. Proyecto de enseñanza

El proyecto de enseñanza es un instrumento que permite evaluar el desempeño del docente o técnico docente a través de una muestra de su práctica. Consiste en la elaboración de un diagnóstico del grupo, una planeación para su puesta en marcha y un texto de análisis que dé cuenta de la reflexión sobre su práctica. Está constituido por tres momentos:

Momento 1. Elaboración del diagnóstico y de la planeación didáctica

Momento 2. Intervención docente

Momento 3. Elaboración de texto de reflexión y análisis de su práctica

Etapa 3. Examen de conocimientos didácticos y curriculares

Este instrumento evalúa los conocimientos didácticos y curriculares que el docente o técnico docente pone en juego para propiciar el aprendizaje de los alumnos, así como sus habilidades para la organización e intervención didáctica. Los principales aspectos a evaluar son los procesos de desarrollo y de aprendizaje infantiles, los propósitos educativos y los contenidos escolares de la Educación Básica, los referentes pedagógicos y los enfoques didácticos del currículo vigente, así como las condiciones para mantener la integridad y seguridad de los alumnos en el aula y en la escuela.

2. 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 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 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 sustentantes con el nivel de precisión requerido para los propósitos de la evaluación, los instrumentos 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 estar organizados jerárquicamente en tres niveles de desagregación: áreas, subáreas y temas, en donde:
  - Cada instrumento debe contar con al menos dos áreas.
  - 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).

Cuestionarios que constituyen la etapa 1:

- En una matriz se deben identificar los indicadores y variables de interés, así como definir sus componentes.
- El contenido debe estar organizado jerárquicamente en dos niveles de desagregación, en donde el primero debe contar, como mínimo, con dos conjuntos de contenidos específicos.

Criterios y parámetros estadísticos

Los instrumentos empleados para la evaluación del desempeño 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:

I. 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.

II. 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 la 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. Estas mediciones deberán aportarse antes de emitir la calificación definitiva de los sustentantes, a fin de salvaguardar la confiabilidad de la decisión.

III. En el caso de los cuestionarios que constituyen la *Etapa 1. Informe de responsabilidades profesionales*, para cada una de las escalas que los constituyen:

- La correlación entre cada reactivo con la puntuación global de la escala deberá ser igual o mayor que 0.20.
- La confiabilidad del constructo medido a través de la escala debe ser igual o mayor que 0.80.

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.

3. Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3

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 la evaluación del desempeño, 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 sustentantes 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 cuatro niveles de desempeño posibles: Nivel I (N I), Nivel II (N II), Nivel III (N III) y Nivel IV (N IV). Los descriptores genéricos para los diferentes grupos de instrumentos y cada nivel se indican en las Tablas 1a y 1b.

Tabla 1a. Descriptores genéricos de los niveles de desempeño para el instrumento Proyecto de enseñanza

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente o técnico docente presenta dificultades para organizar su intervención didáctica, en la que considere las características de sus alumnos, así como las del entorno socio-cultural, escolar y familiar, para el logro del aprendizaje de éstos. Desarrolla su estrategia sin considerar espacios propicios para el aprendizaje, ni la creación de ambientes favorables; en su intervención menciona actividades que realiza con sus alumnos sin vincularlas con los enfoques didácticos de los campos de formación, ni con los aprendizajes esperados; refiere acciones de evaluación que carecen de retroalimentación para la mejora de los aprendizajes de sus alumnos. En la reflexión sobre los resultados de su práctica, presenta dificultades para sustentar sus acciones a partir de los principios filosóficos, normativos y éticos que regulan la profesión docente, así como las estrategias para enriquecer su desarrollo profesional y fortalecer las expectativas que tiene sobre el aprendizaje de sus alumnos.
Nivel II (N II)	El docente o técnico docente considera, en la organización de su intervención didáctica, las características de sus alumnos, así como las del entorno socio-cultural, escolar y familiar, para el logro del aprendizaje de éstos. Desarrolla su estrategia en espacios propicios para el aprendizaje y señala algunos elementos para crear ambientes favorables acordes con los enfoques didácticos de los campos de formación y los aprendizajes esperados. Describe las acciones de evaluación y retroalimentación que dirige a sus alumnos para la mejora de sus aprendizajes. En la reflexión sobre los resultados de su práctica, menciona algunos conceptos filosóficos, normativos y éticos que regulan la profesión docente, así como algunas estrategias para enriquecer su desarrollo profesional y fortalecer las expectativas que tiene sobre el aprendizaje de sus alumnos.
Nivel III (N III)	El docente o técnico docente, en la organización de su intervención didáctica, describe cómo vincula las características de sus alumnos y las del entorno socio-cultural, escolar y familiar con el logro del aprendizaje de éstos. Desarrolla su estrategia en espacios propicios para el aprendizaje en los que logra crear ambientes favorables acordes con los enfoques didácticos de los campos de formación y los aprendizajes esperados. Explica sus acciones de evaluación y de retroalimentación a sus alumnos para la mejora de sus aprendizajes. En la reflexión sobre los resultados de su práctica, explica sus acciones a partir de los principios filosóficos, normativos y éticos que regulan la profesión docente, así como las estrategias que utiliza para enriquecer su desarrollo profesional y fortalecer las expectativas que tiene sobre el aprendizaje de sus alumnos.
Nivel IV (N IV)	El docente o técnico docente argumenta, en la organización de su intervención didáctica, cómo vincula las características de sus alumnos y las del entorno socio-cultural, escolar y familiar con el logro del aprendizaje de éstos. Desarrolla su estrategia en espacios propicios para el aprendizaje en los que logra crear ambientes favorables acordes con los enfoques didácticos de los campos de formación y los aprendizajes esperados. Explica las acciones de evaluación y retroalimentación que dirige a sus alumnos para la mejora de sus aprendizajes. En la reflexión sobre los resultados de su práctica, sustenta sus acciones a partir de los principios filosóficos, normativos y éticos que regulan la profesión docente, ofrece argumentos sobre las estrategias que utiliza para enriquecer su desarrollo profesional y explica cómo va a fortalecer las expectativas que tiene sobre el aprendizaje de sus alumnos.

Tabla 1b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos didácticos y curriculares

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente o técnico docente muestra conocimientos poco consistentes acerca de los procesos de desarrollo y de aprendizaje infantiles, así como de la influencia que tienen los factores familiares, sociales y culturales en sus alumnos. Identifica los propósitos educativos sin reconocer su carácter formativo; demuestra conocimiento de los contenidos escolares de la educación básica, sin lograr establecer su progresión para favorecer el aprendizaje de los alumnos y con ello, el logro de los propósitos educativos. Identifica elementos básicos de los enfoques didácticos del currículo vigente y de sus referentes pedagógicos. Muestra dificultades para identificar estrategias de estudio encaminadas a mejorar su desarrollo profesional.
Nivel II (N II)	El docente o técnico docente reconoce las descripciones de los procesos de desarrollo y de aprendizaje infantiles, así como la influencia que tienen los factores familiares, sociales y culturales en sus alumnos. Identifica los propósitos educativos y reconoce su carácter formativo; demuestra conocimiento de los contenidos escolares de la educación básica y su progresión para favorecer el aprendizaje de los alumnos y con ello, el logro de los propósitos educativos. Reconoce las características de los enfoques didácticos del currículo vigente y sus referentes pedagógicos. Identifica estrategias de estudio encaminadas a mejorar su desarrollo profesional.
Nivel III (N III)	El docente o técnico docente identifica las explicaciones de los procesos de desarrollo y de aprendizaje infantiles, así como la influencia que tienen los factores familiares, sociales y culturales en sus alumnos. Reconoce las características de los propósitos educativos y su carácter formativo; demuestra conocimiento de los contenidos escolares de la educación básica. Señala las explicaciones congruentes con los enfoques didácticos del currículo vigente a partir de referentes pedagógicos. Reconoce estrategias de estudio encaminadas a mejorar su desarrollo profesional.
Nivel IV (N IV)	El docente o técnico docente reconoce el análisis los procesos de desarrollo y de aprendizaje infantiles y argumenta la influencia que tienen los factores familiares, sociales y culturales en sus alumnos. Identifica la explicación del carácter formativo de los propósitos educativos; demuestra conocimiento de los contenidos escolares de la educación básica, su progresión para favorecer el aprendizaje de los alumnos y con ello, el logro de los propósitos educativos. Reconoce las características de los enfoques didácticos del currículo vigente a partir de referentes pedagógicos. Distingue estrategias de estudio encaminadas a mejorar su desarrollo profesional.

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 Instituto determinará el procedimiento a seguir para el establecimiento de los puntos de corte correspondientes, atendiendo al marco jurídico aplicable.

4. Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global

A continuación, se presentan dos subapartados, en el primero se describen los procedimientos para calificar los resultados de los sustentantes en cada instrumento[1] en cada etapa; mientras que en el segundo se detallan los procedimientos para la obtención del resultado global.

4.1 Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen las etapas del proceso de evaluación

4.1.1 Con relación a los instrumentos considerados en las etapas 2 y 3

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

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 sustentante 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 sustentantes (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).

---

[1] En el caso en que el sustentante **no presente alguno** de los instrumentos de evaluación de las etapas 2 y 3 o el cuestionario de autoevaluación de la etapa 1, 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. Para el caso en que el sustentante no presente ninguno de los instrumentos de evaluación de las etapas 2 y 3 ni el cuestionario de autoevaluación de la etapa 1, su resultado global será "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: no presentó", asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos que constituyen el proceso de evaluación del desempeño. En el caso en que la autoridad inmediata no responda el cuestionario que le corresponde de la etapa 1, el resultado en ese instrumento será "SI: sin información".

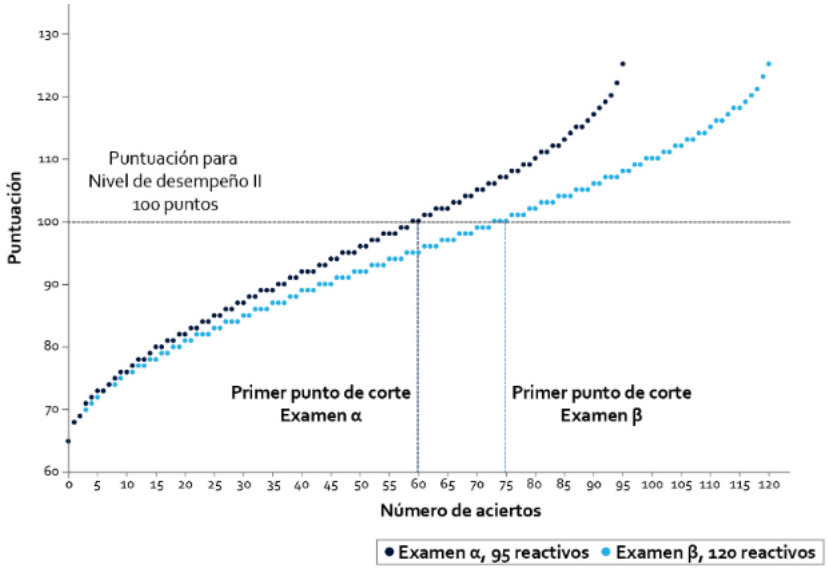
Escala utilizada para reportar los resultados

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 los instrumentos de las etapas 2 y 3 de este proceso de evaluación, permitirá construir una métrica común. 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 los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos[2]).

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 instrumento el punto de corte refiera a número de aciertos diferente. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



4.1.2 Con relación a los cuestionarios que integran la *Etapa 1. Informe de responsabilidades profesionales*

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

[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, los 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; por otra parte, para realizar el escalamiento, el sustentante debe, al menos, haber alcanzado un acierto en el examen; en caso contrario, se reportará como cero y obtendrá N I. Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos, se integrará la información y se definirán cuatro categorías que indicarán el nivel de cumplimiento del sustentante en las responsabilidades profesionales de su función[3]. Cada una de estas categorías tendrá asociada una cantidad de puntos que, como posteriormente se indicará, se adicionará a la puntuación total ponderada, considerando el siguiente orden:

NI: 0 puntos

NII: 1 punto

NIII: 2 puntos

NIV: 3 puntos

Cada cuestionario contribuirá con el 50% de la puntuación de la etapa 1, de tal forma que, en caso de faltar las respuestas de alguno de los dos cuestionarios, la puntuación de la etapa será igual a la puntuación que aporta el cuestionario del que se cuente con información.

En ningún caso, por sí mismo, la omisión de alguno de los dos cuestionarios que considera esta etapa de la evaluación **será causal de un resultado Insuficiente**. Lo anterior porque se trata de reconocer y estimular la participación genuina de los sustentantes y autoridades superiores.

4.2 Resultado global y procedimiento para la conformación de los grupos de desempeño

4.2.1 El resultado global

Para determinar el resultado global de la calificación de los sustentantes, deberán integrarse los resultados de los instrumentos considerados en las tres etapas que conforman el diseño de la evaluación, conforme a los siguientes criterios:

1) Sustentar los dos instrumentos que constituyen las etapas 2 y 3

2) Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3, *el cual debe ser el instrumento de la etapa 3 para el caso de los docentes de inglés en Educación Preescolar, Primaria y Secundaria, y de los docentes de francés en Educación Secundaria*

Cuando no se cumpla con los criterios 1 y 2, no aplicarán los numerales 3, 4 y 5

3) Una vez que se verifica el cumplimiento de los criterios 1 y 2, se calcula la puntuación total ponderada del sustentante, es decir, se pondera[4] el resultado obtenido en los dos instrumentos de las etapas 2 y 3 bajo el siguiente esquema:

a. Etapa 2. Proyecto de enseñanza, 60%

b. Etapa 3. Examen de conocimientos didácticos y curriculares, 40%

4) Se adiciona el resultado obtenido en la etapa 1, de acuerdo con el nivel de cumplimiento alcanzado: NI (0 puntos), NII (1 punto), NIII (2 puntos), o bien NIV (3 puntos).

5) Se asigna el resultado global de la evaluación, que integra los resultados parciales de todo el proceso.

4.2.2 La conformación de los grupos de desempeño

El resultado "Suficiente"

Para alcanzar al menos un resultado suficiente en la evaluación, se deben cumplir los siguientes criterios:

- Sustentar los dos instrumentos que constituyen las etapas 2 y 3
- Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3, *el cual debe ser el instrumento de la etapa 3 para el caso de los docentes de inglés en Educación Preescolar, Primaria y Secundaria, y de los docentes de francés en Educación Secundaria*
- Obtener al menos 100 puntos en la escala de calificación global

Los **grupos de desempeño** estarán conformados únicamente por los sustentantes que obtengan, al menos, un resultado "Suficiente" en la evaluación:

Criterios para formar parte de un grupo de desempeño	
Grupo de desempeño	Puntuación en escala de calificación global
Suficiente	Al menos 100[5] puntos
Bueno	Al menos PC_2G puntos
Destacado	Al menos PC_3G puntos

[3] Para mayores detalles sobre el procedimiento para el escalamiento de las puntuaciones de los cuestionarios, la integración de la información y la asignación de niveles de cumplimiento en la etapa 1, consultar el anexo técnico.

[4] Se traduce como la cantidad de puntos en escala INEE multiplicada por 0.60 y 0.40, respectivamente. Para mayores detalles sobre el algoritmo para el cálculo de la puntuación global, consultar el anexo técnico.

[5] PC_1G siempre es igual a 100, toda vez que el primer punto de corte en los instrumentos considerados en las etapas 2 y 3 siempre es 100. Para mayores detalles sobre el algoritmo para el cálculo de los puntos de corte en la escala de calificación global, consultar el anexo técnico.

El resultado "Insuficiente"

En los siguientes casos se asignará el resultado "Insuficiente" y, por lo tanto, el sustentante **no formará parte de los grupos de desempeño**, **pero recibirá la retroalimentación que corresponda**:

- No sustente los dos instrumentos que constituyen las etapas 2 y 3.
- **No obtenga** al menos NII en por lo menos uno de los dos instrumentos que constituyen las etapas 2 y 3, *además, para el caso de los docentes de inglés en Educación Preescolar, Primaria y Secundaria, y de los docentes de francés en Educación Secundaria, que no obtenga al menos NII en el instrumento de la etapa 3*
- No obtenga **al menos** 100 puntos en la escala de calificación global.

En los dos primeros casos no se dará puntuación global al sustentante.

En los tres casos los sustentantes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan presentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

El resultado "No se presentó a la evaluación"

Para el caso en que el sustentante no presente ninguno de los instrumentos de las etapas 2 y 3 considerados en el diseño de la evaluación, ni el cuestionario de autoevaluación de la etapa 1, en el resultado de la evaluación se indicará: "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: No presentó". Asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos, aun cuando su autoridad inmediata haya respondido el cuestionario que le corresponde de la etapa 1.

Sobre los resultados de la evaluación

El resultado de la evaluación, tanto para los resultados "Insuficientes", como de aquellos que forman parte de un grupo de desempeño ("Suficiente", "Bueno" o "Destacado"), aportará información relevante para diseñar programas y acciones de capacitación, formación y acompañamiento.

Segunda sección.

Evaluación del desempeño en su segunda o tercera oportunidad del personal docente y técnico docente en Educación Básica

De conformidad con la Ley General del Servicio Profesional Docente, esta evaluación del desempeño en su segunda o tercera oportunidad es obligatoria y deberá llevarse a cabo en un plazo no mayor de doce meses después de haberse presentado la primera o segunda evaluación, respectivamente.

Serán sujetos a una segunda o tercera oportunidad de evaluación del desempeño exclusivamente los docentes y técnicos docentes que obtuvieron resultado insuficiente en su primera o segunda evaluación del desempeño, respectivamente.

La calificación global se estimará siguiendo el mismo modelo de calificación desarrollado en los presentes criterios técnicos (véase la primera sección). Se considerarán los resultados obtenidos en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) con base en las siguientes equivalencias:

Equivalencias para la etapa 2

Se recuperará la información de los resultados que el sustentante haya obtenido en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) en los siguientes instrumentos de evaluación:

- *Planeación didáctica argumentada*
- *Expediente de evidencias de enseñanza*

Las reglas de equivalencias serán las siguientes:

Resultado obtenido en Planeación didáctica argumentada	***Resultado obtenido en Expediente de evidencias de enseñanza***	**Resultado asignado para la etapa 2**
NII, NIII o NIV	*NII, NIII o NIV*	El nivel de desempeño más alto que haya alcanzado en cualquiera de los dos instrumentos
En cualquier resultado cuya combinación de los dos instrumentos sea: NP o NI con NP, NI, NII, NIII o NIV		Debe presentar el instrumento de la etapa

Equivalencias para la etapa 3

Se recuperará la información de los resultados que el sustentante haya obtenido en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) en los siguientes instrumentos de evaluación:

- *Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos*
- *Examen complementario o Examen de dominio de la lengua indígena (para los casos en que aplique)*

Las reglas de equivalencias para **docentes y técnicos docentes cuya anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) no consideraba un examen complementario o de dominio de la lengua indígena** serán las siguientes:

Resultado obtenido en Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos	**Resultado asignado para la etapa 3**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NI o NP	Debe presentar el instrumento de la etapa

Para los **docentes cuya anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) consideraba un examen complementario o un examen de dominio de la lengua indígena,** se consideran para esta evaluación dos instrumentos en la Etapa 3 (Examen de conocimientos didácticos y curriculares y examen complementario o examen de dominio de la lengua indígena), por lo que las equivalencias serán las siguientes:

Resultado obtenido en Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos	**Resultado asignado al Examen de conocimientos didácticos y curriculares de la etapa 3**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NP o NI	Debe presentar el Examen de conocimientos didácticos y curriculares

Resultado obtenido en Examen complementario o Examen de dominio de la lengua indígena	**Resultado asignado al Instrumento que le corresponda de la etapa 3 (Examen complementario o Examen de dominio de la lengua indígena)**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NP o NI	Debe presentar el instrumento que le corresponda de la etapa (Examen complementario o Examen de dominio de la lengua indígena)

Para estos docentes cuya anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) consideraba un examen complementario o un examen de dominio de la lengua indígena, el resultado global se determinará integrando los resultados de los instrumentos considerados en las tres etapas que conforman el diseño de la evaluación, conforme a los siguientes criterios:

1) Sustentar los tres instrumentos que constituyen las etapas 2 y 3

2) Obtener al menos NII en por lo menos dos de los tres instrumentos de las etapas 2 y 3, *uno de los cuales debe ser el Examen de dominio de la lengua indígena para el caso de los docentes de lengua indígena*.

Cuando no se cumpla con los criterios 1 y 2, no aplicarán los numerales 3, 4 y 5

3) Una vez que se verifica el cumplimiento de los criterios 1 y 2, se calcula la puntuación total ponderada del sustentante, es decir, se pondera[6] el resultado obtenido en los tres instrumentos de las etapas 2 y 3 bajo el siguiente esquema:

a. Etapa 2. Proyecto de enseñanza, 60%

b. Etapa 3. Examen de conocimientos didácticos y curriculares, 40%

- Examen de conocimientos didácticos y curriculares, 20%
- Examen complementario o Examen de dominio de la lengua indígena, 20%

4) Se adiciona el resultado obtenido en la etapa 1, de acuerdo con el nivel de cumplimiento alcanzado: NI (0 puntos), NII (1 punto), NIII (2 puntos), o bien NIV (3 puntos)

5) Se asigna el resultado global de la evaluación, que integra los resultados parciales de todo el proceso

De esta forma, para alcanzar al menos un **resultado Suficiente** en la evaluación, estos sustentantes deben cumplir los siguientes criterios:

- Sustentar los tres instrumentos que constituyen las etapas 2 y 3
- Obtener al menos NII en por lo menos dos de los tres instrumentos de las etapas 2 y 3, *uno de los cuales debe ser el Examen de dominio de la lengua indígena para el caso de los docentes de lengua indígena*
- Obtener al menos 100 puntos en la escala de calificación global

Asimismo, en los siguientes casos se asignará el resultado **Insuficiente** y, por lo tanto, el sustentante **no formará parte de los grupos de desempeño**, **pero recibirá la retroalimentación que corresponda**:

- No sustente los tres instrumentos que constituyen las etapas 2 y 3.
- **No obtenga** al menos NII en por lo menos dos de los tres instrumentos que constituyen las etapas 2 y 3, *además, para el caso de los docentes de lengua indígena, que no obtenga al menos NII en el examen de dominio de la lengua indígena*
- No obtenga **al menos** 100 puntos en la escala de calificación global.

En los dos primeros casos no se dará puntuación global al sustentante.

En los tres casos los sustentantes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan presentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

Finalmente, cualquier situación no prevista en los presentes criterios técnicos será analizada por la Junta de Gobierno 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* y toda vez que los cuestionarios que constituyen la etapa 1 de este proceso tienen como finalidad recabar información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función profesional, y **únicamente** pueden ser considerados para **adicionar puntos al sustentante en su calificación global, la cual está en función de los resultados alcanzados en los instrumentos que constituyen las etapas 2 y 3**, es fundamental señalar que, en ningún caso, **se puede considerar solamente un instrumento** para integrar la calificación de los sustentantes conforme al diseño de la evaluación, es decir:

> *Ninguna decisión que tenga consecuencias importantes sobre los individuos o instituciones, se basará únicamente en los resultados de sólo un instrumento de evaluación, por lo cual, deberán considerarse otras fuentes confiables de información que incrementen la validez de las decisiones que se tomen.*

Lo anterior debido a que la evidencia empírica que resulte del análisis psicométrico de los instrumentos de la segunda y tercera etapa de la evaluación del desempeño del personal docente y técnico docente debe mostrar que, una vez que éstos fueron aplicados, cumplen con los criterios técnicos establecidos por el Instituto, de esta forma la integración de los resultados de la evaluación debe permitir establecer inferencias válidas sobre el desempeño y competencias de los sustentantes evaluados.

[6] Se traduce como la cantidad de puntos en escala INEE multiplicada por 0.60, 0.20 y 0.20, respectivamente. La puntuación de la Etapa 3 se calcula considerando que cada uno de los dos exámenes que la componen aporta el 50%. Para mayores detalles sobre el algoritmo para el cálculo de la puntuación global, consultar el anexo técnico.

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 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 ni el centro de trabajo 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 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 calificación que se asigna 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 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 del sustentante en cada aspecto a evaluar para emitir su calificación definitiva 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**, toda vez que la administración de éstos es *in situ*, si existen discrepancias entre los jueces en cuanto a la asignación de categorías en algunos aspectos a evaluar, se les pedirá que lleguen a un acuerdo, a fin de establecer la categoría definitiva para el sustentante en dichos aspectos a evaluar.

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 del punto 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 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:

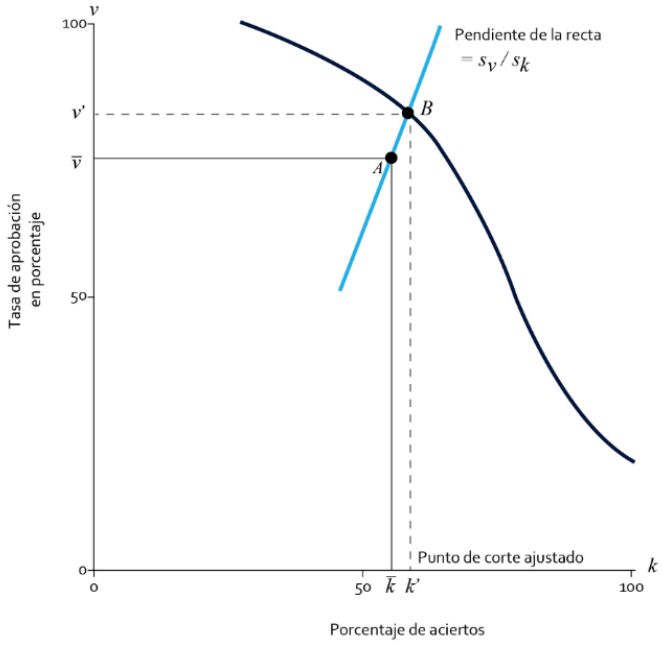
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas ($\bar{k}$, $\bar{v}$), (ver siguiente figura).

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 de los instrumentos considerados en las etapas 2 y 3

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 II) 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).

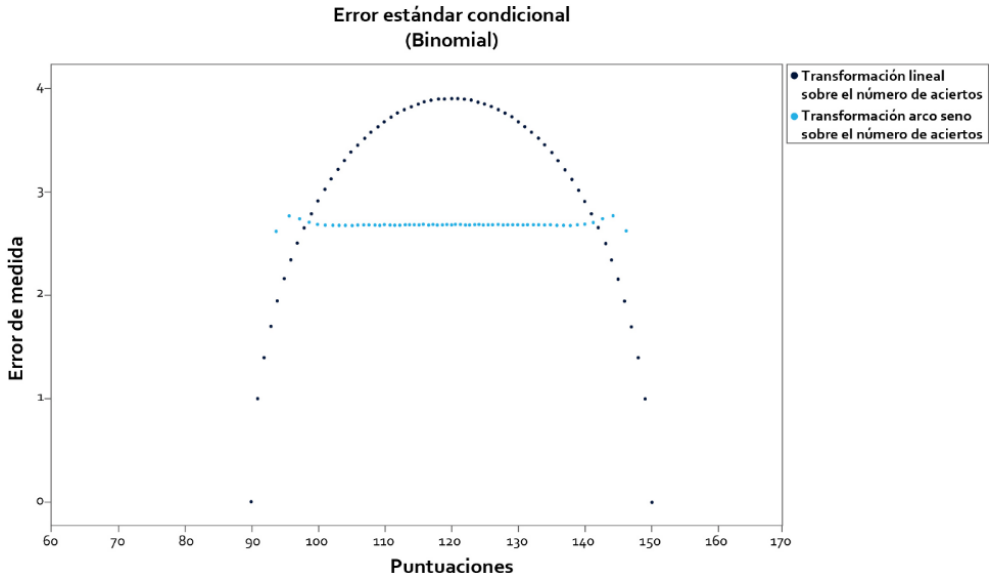
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

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 N I.

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

Como se comentó, 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 II (N II) en los instrumentos de evaluación, se ubicará en el valor 100. Para determinarla se empleará la siguiente ecuación:

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

Donde $\boldsymbol{A = \frac{Q}{[c(K) - c(0)]}}$, $\boldsymbol{B = 100 - A * c(PC1)}$, *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 II.

El valor de Q dependerá 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 (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 puntuaciones 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 1 (P_{A1i}), 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 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, 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 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, el punto 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.

$\boldsymbol{\alpha}$ 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:

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

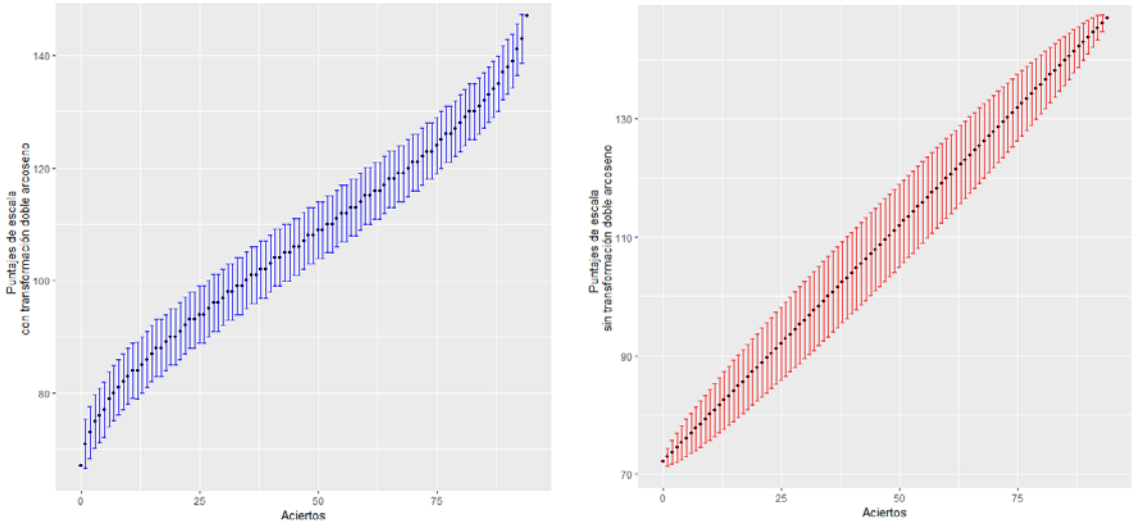
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

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.

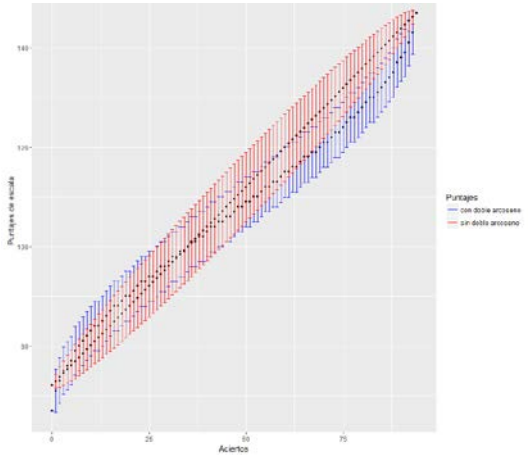
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 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 II (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

Como ya se indicó en el cuerpo del documento, 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:

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

Donde s denota la población sintética y

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

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

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

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

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

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

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.

Es importante señalar que para los puntajes que se les aplique la equiparación $x_e = b_1x + b_0$, 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:

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

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).

Procedimiento para el escalamiento de las puntuaciones de los cuestionarios de la etapa 1

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos.

La escala de puntuaciones de cada cuestionario se ubicará en el intervalo [0, 50], si un cuestionario no es presentado se le asignará una puntuación de cero. Ambos cuestionarios serán escalados utilizando el modelo de crédito parcial. Para que el rango de puntuaciones vaya de 0 a 50, las puntuaciones que se obtengan con el modelo se escalarán linealmente y se redondearán al entero más próximo, utilizando el criterio de que puntuaciones con cinco décimas o más, suben al siguiente entero.

De esta forma, la puntuación alcanzada en la etapa 1 será calculada como la suma de las puntuaciones de ambos cuestionarios, por lo que se ubicará en el intervalo [0, 100].

La asignación del nivel de cumplimiento en la etapa 1 y la cantidad de puntos que se adicionan a la puntuación total del sustentante, será con base en la siguiente tabla:

Suma de las puntuaciones de ambos cuestionarios	Nivel de cumplimiento	Puntos que se adicionan
De 0 a 25	NI	0
De 26 a 50	NII	1
De 51 a 75	NIII	2
De 76 a 100	NIV	3

Algoritmo para el cálculo de la puntuación global

Una vez que se ha verificado que el sustentante presentó los dos instrumentos que constituyen las etapas 2 y 3 del proceso de evaluación y que obtuvo al menos NII en por lo menos uno de ellos, se procede a calcular la puntuación global con base en el siguiente esquema:

Etapa 2. Proyecto de enseñanza, 60%

Etapa 3. Examen de conocimientos didácticos y curriculares, 40%

$$G_i = 0.60 * P_{1i} + 0.40 * P_{2i} + P_{Ei}$$

$G_i = Puntuación\ global\ que\ alcanza\ el\ sustentante\ i\ en\ la\ evaluación$

$P_{1i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Proyecto\ de\ enseñanza$
$P_{2i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ de\ conocimientos\ didácticos\ y\ curriculares$
$P_{Ei} = 0,1,2,3\ (Puntuación\ que\ se\ adiciona\ con\ base\ en\ el\ resultado\ del\ sustentante\ i\ en\ la\ etapa\ 1)$

Para el caso de docentes cuya anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) consideraba un examen complementario o un examen de dominio de la lengua indígena, una vez que se ha verificado que presentó los tres instrumentos que constituyen las etapas 2 y 3 del proceso de evaluación y que obtuvo al menos NII en por lo menos dos de ellos (uno de los cuales debe ser el Examen de dominio de la lengua indígena para el caso de los docentes de lengua indígena), se procede a calcular la puntuación global con base en el siguiente esquema:

Etapa 2. Proyecto de enseñanza, 60%

Etapa 3. Examen de conocimientos didácticos y curriculares, 40%

- Examen de conocimientos didácticos y curriculares, 20%
- Examen complementario o Examen de dominio de la lengua indígena, 20%

$$G_i = 0.60 * P_{1i} + 0.20 * P_{2i} + 0.20 * P_{3i} + P_{Ei}$$

$G_i = Puntuación\ global\ que\ alcanza\ el\ sustentante\ i\ en\ la\ evaluación$

$P_{1i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Proyecto\ de\ enseñanza$

$P_{2i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ de\ conocimientosdidácticos\ y\ curriculares$

$P_{3i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ complementario\ o\ Examen\ de\ dominio\ de\ la\ lengua\ indígena\ (según\ corresponda)$

$P_{Ei} = 0,1,2,3\ (Puntuación\ que\ se\ adiciona\ con\ base\ en\ el\ resultado\ del\ sustentante\ i\ en\ la\ etapa\ 1)$

Algoritmo para el cálculo de los puntos de corte en la escala de calificación global

Los puntos de corte en la escala global se calcularán considerando los puntos de corte establecidos en los instrumentos utilizados en las etapas 2 y 3, con base en el siguiente algoritmo:

$$PC_iG = 0.60 * PC_iP + 0.40 * PC_iE$$

$i = 1, 2, 3$

$PC_iG = Punto\ de\ corte\ i\ en\ la\ escala\ de\ calificación\ global$

$PC_iP = Punto\ de\ corte\ i\ establecido\ en\ el\ Proyecto\ de\ enseñanza$

$PC_iE = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ de\ conocimientos\ didácticos\ y\ curriculares$

Para el caso de docentes cuya anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) consideraba un examen complementario o un examen de dominio de la lengua indígena, se utilizará el siguiente algoritmo:

$$PC_iG = 0.60 * PC_iP + 0.20 * PC_iE + 0.20 * PC_iEC$$

$i = 1, 2, 3$

$PC_iG = Punto\ de\ corte\ i\ en\ la\ escala\ de\ calificación\ global$

$PC_iP = Punto\ de\ corte\ i\ establecido\ en\ el\ Proyecto\ de\ enseñanza$

$PC_iE = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ de\ conocimientos\ didácticos\ y\ curriculares$

$PC_iEC = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ complementario\ o\ Examen\ de\ dominio\ de\ la\ lengua\ indígena\ (según\ corresponda)$

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.

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.

Bentler, P. M. (2006). EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

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.

Brennan, R. L. (2012). Scaling PARCC Assessments: Some considerations and a synthetic data example en: http://parcconline.org/about/leadership/12-technical-advisory-committee

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

Downing, SM (2004). Reliability: On the reproducibility of assessment data. Med Educ; 38(9): 1006-1012. 21

Holland, P. W., & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), Statistical models for test equating, scaling, and linking (pp. 89–107). New York, NY: Springer

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. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Masters, Geoff (1982). A Rasch model for Partial Credit Scoring. Psychometrika-vol. 47, No. 2.

Muñiz, José (2003): Teoría clásica de los test. Ediciones pirámide, Madrid.

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

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

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

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

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

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

Shun-Wen Chang (2006) Methods in Scaling the Basic Competence Test, *Educational and Psychological Measurement,* 66 (6) 907-927.

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

Stemler, E. & Tsai, J. (2008). *Best Practices in Interrater Reliability Three Common Approaches* in Best practices in quantitative methods (pp. 29–49). SAGE Publications, Inc.

Thompson, Bruce ed. (2003): Score reliability. Contemporary thinking on reliability issues. SAGE Publications, Inc.

Wilson, Mark (2005). Constructing measures. An ítem response modeling approach. Lawrence Erlbaum Associates, Publishers.

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.

Wu, Margaret & Adams, Ray (2007). Applying the Rasch Model to Psycho-social measurement. A practical approach. Educational measurement solutions, Melbourne.

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

Ciudad de México, a veintiocho de septiembre de dos mil diecisiete.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil diecisiete, celebrada el veintiocho de septiembre de dos mil diecisiete. Acuerdo número **SOJG/09-17/05,R.** El Consejero Presidente, **Eduardo Backhoff Escudero**.- Rúbrica.- Los Consejeros: **Gilberto Ramón Guevara Niebla, Sylvia Irene Schmelkes del Valle, Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 457556)

CRITERIOS técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Media Superior en el ciclo escolar 2017-2018.

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 EMISIÓN DE RESULTADOS PARA LLEVAR A CABO LA EVALUACIÓN DEL DESEMPEÑO DEL PERSONAL DOCENTE Y TÉCNICO DOCENTE EN EDUCACIÓN MEDIA SUPERIOR EN EL CICLO ESCOLAR 2017-2018.

El presente documento está dirigido a las autoridades educativas que en el marco de sus atribuciones implementan evaluaciones que, por la naturaleza de sus resultados, regula el Instituto Nacional para la Evaluación de la Educación (INEE), en especial las referidas al Servicio Profesional Docente (SPD) que son desarrolladas por la Coordinación Nacional del Servicio Profesional Docente (CNSPD).

Con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 7, fracción X de la Ley General del Servicio Profesional Docente; 22, 28, fracción X, 38, fracciones VI, IX y XXII de la Ley del Instituto Nacional para la Evaluación de la Educación; en los Lineamientos para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Básica y Media Superior en el ciclo escolar 2017-2018 (LINEE-04-2017), la Junta de Gobierno aprueba los siguientes criterios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados para llevar a cabo la evaluación del desempeño del personal docente y técnico docente en Educación Media Superior en el ciclo escolar 2017-2018.

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. Su contenido se organiza de la siguiente manera:

Primera sección: Sobre la evaluación del desempeño para el ciclo escolar 2017-2018.

Incorpora cinco apartados: 1) Características generales de los instrumentos para evaluar el desempeño del personal docente y técnico docente; 2) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 3) Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3; 4) Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global.

Segunda sección: Sobre la evaluación del desempeño de quienes será su segunda o tercera oportunidad.

En la parte final se presenta un Anexo técnico con información detallada de algunos de los aspectos técnicos que se consideran en el documento.

Definición de términos

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

I. **Alto impacto:** Se indica cuando los resultados de un 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.

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. **Cuestionario:** Tipo de instrumento de evaluación que sirve para recolectar información sobre actitudes, conductas, opiniones, contextos demográficos o socioculturales, entre otros.

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

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

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 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 una misma persona en un mismo instrumento que tienden a distribuirse alrededor de un puntaje verdadero. El puntaje verdadero siempre es desconocido porque ninguna medida puede ser una representación perfecta de un puntaje verdadero.

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. Constituye el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XIX. **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.

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

XXI. **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.

XXII. **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.

XXIII. **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.

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

XXV. **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.

XXVI. **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.

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

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

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

XXX. **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.

XXXI. **Porcentaje de acuerdos intra-jueces:** Medida del grado en que el mismo juez, a través de dos o más mediciones repetidas a los mismos sujetos que evalúa, coincide en la puntuación asignada al desempeño de los sujetos, evaluados a través de una rúbrica.

XXXII. **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.

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

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

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

XXXVI. **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.

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

XXXVIII. **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.

Primera sección.

Evaluación del desempeño del personal docente y técnico docente en Educación Media Superior, 2017-2018

1. Características generales de los instrumentos para evaluar el desempeño del personal docente y técnico docente

La evaluación del desempeño es un proceso integrado que incluye varios instrumentos que dan cuenta de los diferentes aspectos que se describen en los Perfiles, parámetros e indicadores establecidos por la autoridad educativa. A continuación, se describen sucintamente cada uno de los instrumentos considerados en cada etapa del proceso.

Etapa 1. Informe de responsabilidades profesionales

Esta etapa está constituida por dos instrumentos de evaluación, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función del personal docente y técnico docente, así como identificar las fortalezas y aspectos a mejorar en su práctica:

a) Cuestionario de autoevaluación, respondido por el docente o técnico docente

b) Cuestionario para su autoridad inmediata, quien proporcionará la información relativa al nivel de cumplimiento de las responsabilidades profesionales del docente o técnico docente

Etapa 2. Proyecto de enseñanza

El proyecto de enseñanza es un instrumento que permite evaluar el desempeño del docente o técnico docente a través de una muestra de su práctica. Consiste en la elaboración de un diagnóstico del grupo, una planeación para su puesta en marcha y un texto de análisis que dé cuenta de la reflexión sobre su práctica. Está constituido por tres momentos:

Momento 1. Elaboración del diagnóstico y de la secuencia didáctica

Momento 2. Intervención docente

Momento 3. Elaboración de texto de reflexión y análisis de su práctica

Etapa 3. Examen de conocimientos y habilidades didácticas

Esta etapa se divide en dos exámenes: el examen de conocimientos y el examen de habilidades didácticas.

El examen de conocimientos evalúa el manejo de conocimientos disciplinares por parte del docente y de conocimientos científicos y tecnológicos en el caso del técnico docente, que favorecen el aprendizaje de los estudiantes; el examen de habilidades didácticas tiene como propósito identificar los conocimientos y las habilidades del docente y técnico docente para propiciar la mejora de sus prácticas y fortalecer su función.

Docentes	Técnicos docentes
Examen de conocimientos disciplinares	Examen de conocimientos científicos y tecnológicos
Examen de habilidades didácticas	Examen de habilidades didácticas

2. 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 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 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 sustentantes con el nivel de precisión requerido para los propósitos de la evaluación, los instrumentos 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 estar organizados jerárquicamente en tres niveles de desagregación: áreas, subáreas y temas, en donde:
  - Cada instrumento debe contar con al menos dos áreas.
  - 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).

Cuestionarios que constituyen la etapa 1:

- En una matriz se deben identificar los indicadores y variables de interés, así como definir sus componentes.
- El contenido debe estar organizado jerárquicamente en dos niveles de desagregación, en donde el primero debe contar, como mínimo, con dos conjuntos de contenidos específicos.

Criterios y parámetros estadísticos

Los instrumentos empleados para la evaluación del desempeño 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:

I. 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.

II. 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 la 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. Estas mediciones deberán aportarse antes de emitir la calificación definitiva de los sustentantes, a fin de salvaguardar la confiabilidad de la decisión.

III. En el caso de los cuestionarios que constituyen la *Etapa 1. Informe de responsabilidades profesionales*, para cada una de las escalas que los constituyen:

- La correlación entre cada reactivo con la puntuación global de la escala deberá ser igual o mayor que 0.20.
- La confiabilidad del constructo medido a través de la escala debe ser igual o mayor que 0.80.

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.

3. Procedimiento para el establecimiento de puntos de corte y estándares de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3

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 la evaluación del desempeño, 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 sustentantes 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 cuatro niveles de desempeño posibles: Nivel I (N I), Nivel II (N II), Nivel III (N III) y Nivel IV (N IV). Los descriptores genéricos para los diferentes grupos de instrumentos y cada nivel se indican en las Tablas 1a, 1b, 1c y 1d.

Tabla 1a. Descriptores genéricos de los niveles de desempeño para el instrumento Proyecto de enseñanza

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente o técnico docente demuestra conocimientos poco consistentes acerca de los procesos de construcción del conocimiento y del aprendizaje, así como de los procesos de enseñanza basados en el modelo por competencias; tiene dificultades para explicar la naturaleza, métodos y consistencia lógica de los saberes de la asignatura o módulo que imparte y para identificar las características y necesidades de aprendizaje de los estudiantes. En su planeación esboza el desarrollo de competencias, sin considerar los conocimientos previos de los estudiantes, ni estrategias y técnicas vinculadas al contexto de los mismos. En el desarrollo de su práctica docente recurre de manera elemental a las tecnologías de la información y la comunicación para el logro de los aprendizajes. Refiere estrategias para la evaluación de los aprendizajes, sin establecer el uso de los resultados para retroalimentar a los estudiantes y su propia práctica.
Nivel II (N II)	El docente o técnico docente describe los procesos de construcción del conocimiento y del aprendizaje, así como los procesos de enseñanza basados en el modelo por competencias; describe la naturaleza, métodos y consistencia lógica de los saberes de la asignatura o módulo que imparte, así como las características y necesidades de aprendizaje de los estudiantes. En su planeación considera los conocimientos previos de los estudiantes, así como estrategias y técnicas vinculadas al contexto de los mismos, para el desarrollo de competencias. En su práctica docente, identifica algunas herramientas de las tecnologías de la información y la comunicación para el logro de los aprendizajes. Describe estrategias para la evaluación de los aprendizajes, así como el uso de los resultados para retroalimentar a los estudiantes y su propia práctica.
Nivel III (N III)	El docente o técnico docente explica los procesos de construcción del conocimiento y del aprendizaje, así como de los procesos de enseñanza basados en el modelo por competencias; Explica la naturaleza, métodos y consistencia lógica de los saberes de la asignatura o módulo que imparte, así como las características y necesidades de aprendizaje de los estudiantes. En su planeación considera los conocimientos previos de los estudiantes, así como estrategias y técnicas vinculadas al contexto de los mismos, para el desarrollo de competencias. En su práctica docente, recurre a algunas herramientas de las tecnologías de la información y la comunicación para el logro de los aprendizajes. Explica estrategias para la evaluación de los aprendizajes, así como el uso de los resultados para retroalimentar a los estudiantes y su propia práctica.
Nivel IV (N IV)	El docente o técnico docente analiza los procesos de construcción del conocimiento y del aprendizaje para justificar los procesos de enseñanza basados en el modelo por competencias. Analiza la naturaleza, métodos y consistencia lógica de los saberes de la asignatura o módulo que imparte para vincularlos con las características y necesidades de aprendizaje de los estudiantes. En su planeación considera los conocimientos previos de los estudiantes, así como estrategias y técnicas vinculadas al contexto de los mismos, para el desarrollo de competencias. En su práctica docente, utiliza herramientas diversas de las tecnologías de la información y la comunicación para el logro de los aprendizajes. Utiliza diversas estrategias para la evaluación de los aprendizajes y retroalimenta a sus estudiantes y su propia práctica a partir de sus resultados.

Tabla 1b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos disciplinares

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente demuestra conocimientos poco consistentes para organizar el contenido teórico-metodológico para el logro de los propósitos de la asignatura que imparte, así como para seleccionar estrategias para el proceso de enseñanza y aprendizaje relacionadas con la asignatura que imparte. Tiene dificultades para identificar acciones para promover en los estudiantes el pensamiento inductivo y deductivo para el análisis o solución de problemas cotidianos y relacionados con su asignatura, así como para distinguir metodologías para investigar y proponer soluciones en situaciones que se relacionan con el campo de su asignatura.
Nivel II (N II)	El docente demuestra conocimientos básicos para organizar el contenido teórico-metodológico de la asignatura que imparte, para el logro de los propósitos educativos, selecciona algunas estrategias de enseñanza y aprendizaje que no se relacionan estrechamente con la asignatura que imparte. Reconoce algunas acciones para promover en los estudiantes el pensamiento inductivo y deductivo para el análisis o solución de problemas cotidianos y relacionados con su asignatura. Conoce algunas metodologías para investigar y proponer soluciones en situaciones que se relacionan con el campo de su asignatura.
Nivel III (N III)	El docente demuestra conocimientos sólidos para organizar el contenido teórico-metodológico para el logro de los propósitos de la asignatura que imparte, selecciona estrategias para el proceso de enseñanza y aprendizaje de la asignatura que imparte y reconoce acciones para promover en los estudiantes el pensamiento inductivo y deductivo y para el análisis o solución de problemas cotidianos, relacionados con su asignatura. Conoce metodologías para investigar y proponer soluciones fundamentadas en situaciones que se relacionan con el campo de su asignatura.
Nivel IV (N IV)	El docente demuestra conocimientos y habilidades sólidos para organizar el contenido teórico-metodológico para el logro de los propósitos de la asignatura que imparte, selecciona estrategias adecuadas para el proceso de enseñanza y aprendizaje de la asignatura que imparte y reconoce la importancia de promover en los estudiantes el pensamiento inductivo y deductivo para el análisis o solución de problemas cotidianos y relacionados con su asignatura. Utiliza diversas metodologías para investigar, fundamentar y proponer soluciones a situaciones que se relacionan con el campo de su asignatura.

Tabla 1c. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos científicos y tecnológicos

Nivel de desempeño	Descriptor
Nivel I (N I)	El técnico docente demuestra carencia de conocimientos en la aplicación de estrategias de enseñanza y aprendizaje con base en la transversalidad entre las asignaturas del plan de estudios correspondiente, desconoce las distintas formas de organización del contenido teórico-metodológico para el logro de los propósitos de las asignaturas correspondientes a su campo disciplinar. Asimismo, selecciona estrategias poco adecuadas para el proceso de enseñanza y aprendizaje y emplea de forma limitada las tecnologías de la información y la comunicación que se encuentran disponibles en su contexto como herramientas para su práctica docente o para favorecer los aprendizajes de los alumnos.
Nivel II (N II)	El técnico docente demuestra conocimientos básicos en la aplicación de estrategias de enseñanza y aprendizaje con base en la transversalidad entre las asignaturas del plan de estudios correspondiente, además reconoce algunas formas de organización del contenido teórico-metodológico para el logro parcial de los propósitos de las asignaturas correspondientes a su campo disciplinar. Asimismo, selecciona algunas estrategias para el proceso de enseñanza y aprendizaje pero que carecen de vínculo con su asignatura, emplea escasamente las tecnologías de la información y la comunicación que se encuentran disponibles en su contexto como herramientas para su práctica docente o para favorecer los aprendizajes de los alumnos.

Nivel III (N III)	El técnico docente demuestra conocimientos sólidos en la aplicación de estrategias de enseñanza y aprendizaje con base en la transversalidad entre las asignaturas del plan de estudios correspondiente, además reconoce algunas formas de organización del contenido teórico-metodológico para el logro parcial de los propósitos de las asignaturas correspondientes a su campo disciplinar. Asimismo, selecciona estrategias para el proceso de enseñanza y aprendizaje aterrizadas en su asignatura, emplea como herramientas para su práctica docente y para favorecer los aprendizajes de los alumnos las tecnologías de la información y la comunicación que se encuentran disponibles en su contexto.
Nivel IV (N IV)	El técnico docente demuestra conocimientos y habilidades sólidos en la aplicación de estrategias de enseñanza y aprendizaje con base en la transversalidad entre las asignaturas del plan de estudios correspondiente, además reconoce las distintas formas de organización del contenido teórico-metodológico para el logro de los propósitos de las asignaturas correspondientes a su campo disciplinar. Asimismo, selecciona estrategias adecuadas para el proceso de enseñanza y aprendizaje aterrizadas en su asignatura, emplea como herramientas para su práctica docente y para favorecer los aprendizajes de los alumnos las tecnologías de la información y la comunicación que se encuentran disponibles en su contexto.

Tabla 1d. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de habilidades didácticas

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente o técnico docente demuestra carencia de conocimientos sobre los procesos de construcción del conocimiento, enseñanza y aprendizaje, muestra dificultades para identificar características y necesidades de aprendizaje de los estudiantes para su formación académica y para seleccionar estrategias de evaluación y retroalimentación de los aprendizajes para el desarrollo de los procesos de formación de los estudiantes. Distingue planes de trabajo que incorporan estrategias y técnicas orientadas al desarrollo de competencias, pero que carecen de vinculación con el contexto de los estudiantes, adicionalmente, reconoce la posibilidad del uso de tecnologías de la información y de la comunicación, pero carece de relación con su práctica docente. Desconoce el vínculo entre el entorno sociocultural y escolar, así como de los intereses de los estudiantes con su práctica docente. Reconoce las disposiciones legales e institucionales, pero muestra dificultad para atenderlas en su práctica docente.
Nivel II (N II)	El docente o técnico docente demuestra conocimientos básicos en los procesos de construcción del conocimiento, enseñanza y aprendizaje basados en el modelo por competencias, identifica características y necesidades de aprendizaje de los estudiantes para su formación académica pero selecciona estrategias poco adecuadas de evaluación y retroalimentación de los aprendizajes para el desarrollo de los procesos de formación de los estudiantes, reconociendo ocasionalmente partes del marco normativo vigente. Identifica planes de trabajo que incorporan estrategias y técnicas orientadas al desarrollo de competencias, pero que carecen de vinculación con el contexto de los estudiantes; asimismo, relaciona escasamente su práctica docente con el uso de tecnologías de la información y de la comunicación, el entorno sociocultural y escolar y los intereses de los estudiantes. Reconoce de forma limitada la importancia de establecer ambientes éticos, incluyentes y equitativos entre los estudiantes y atiende parcialmente las disposiciones legales e institucionales en su práctica docente.
Nivel III (N III)	El docente o técnico docente demuestra conocimientos sólidos en los procesos de construcción del conocimiento, enseñanza y aprendizaje basados en el modelo por competencias, identifica algunas características y necesidades de aprendizaje de los estudiantes para su formación académica y selecciona estrategias de evaluación y retroalimentación de los aprendizajes para el desarrollo de los procesos de formación de los estudiantes, reconociendo ocasionalmente el marco normativo vigente. Identifica planes de trabajo que incorporan, de manera general, estrategias y técnicas orientadas al desarrollo de competencias que se vinculan con el contexto de los estudiantes, adicionalmente, reconoce la posibilidad del uso de tecnologías de la información y de la comunicación como herramientas de su práctica docente. Relaciona ocasionalmente el entorno sociocultural y escolar, así como los intereses de los estudiantes con su práctica docente. Reconoce la importancia de establecer ambientes éticos, incluyentes y equitativos entre los estudiantes y atiende parcialmente las disposiciones legales e institucionales en su práctica docente.

Nivel IV (N IV)	El docente o técnico docente demuestra conocimientos y habilidades sólidos en los procesos de construcción del conocimiento, enseñanza y aprendizaje basados en el modelo por competencias, identifica plenamente las características y necesidades de aprendizaje de los estudiantes para su formación académica y selecciona adecuadamente estrategias de evaluación y retroalimentación de los aprendizajes para el desarrollo de los procesos de formación de los estudiantes, reconociendo el marco normativo vigente. Identifica los planes de trabajo que incorporan estrategias y técnicas orientadas al desarrollo de competencias que se vinculan con el contexto de los estudiantes, adicionalmente, reconoce la importancia del uso de tecnologías de la información y de la comunicación como herramientas de su práctica docente. Relaciona constantemente el entorno sociocultural y escolar, así como los intereses de los estudiantes con su práctica docente. Reconoce la importancia de establecer ambientes éticos, incluyentes y equitativos entre los estudiantes, además atiende las disposiciones legales e institucionales en su práctica docente.

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 Instituto determinará el procedimiento a seguir para el establecimiento de los puntos de corte correspondientes, atendiendo al marco jurídico aplicable.

4. Resultado de la evaluación del desempeño: resultado por etapa e instrumento y resultado global

A continuación, se presentan dos subapartados, en el primero se describen los procedimientos para calificar los resultados de los sustentantes en cada instrumento[1] en cada etapa; mientras que en el segundo se detallan los procedimientos para la obtención del resultado global.

4.1 Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen las etapas del proceso de evaluación

4.1.1 Con relación a los instrumentos considerados en las etapas 2 y 3

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

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 sustentante 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 sustentantes (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).

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 los instrumentos de las etapas 2 y 3 de este proceso de evaluación, permitirá construir una métrica común. 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 los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos[2]).

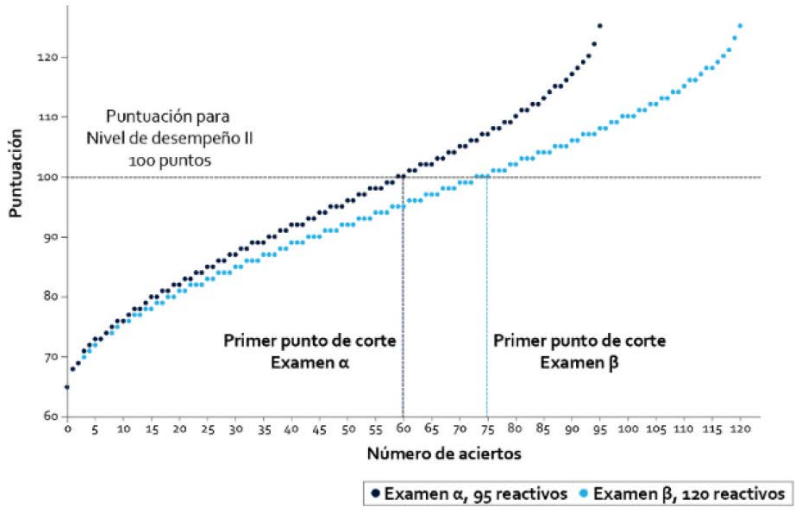
Al utilizar esta escala, diferente a las escalas que se utilizan para reportar resultados de aprendizaje en el aula (de 5 a 10 o de 0% a 100%, donde el 6 o 60% de aciertos es aprobatorio), se evita que se realicen interpretaciones equivocadas de los resultados obtenidos en los exámenes, en virtud de 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.

---

[1] En el caso en que el sustentante **no presente alguno** de los instrumentos de evaluación de las etapas 2 y 3 o el cuestionario de autoevaluación de la etapa 1, 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. Para el caso en que el sustentante no presente ninguno de los instrumentos de evaluación de las etapas 2 y 3 ni el cuestionario de autoevaluación de la etapa 1, su resultado global será "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: no presentó", asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos que constituyen el proceso de evaluación del desempeño. En el caso en que la autoridad inmediata no responda el cuestionario que le corresponde de la etapa 1, el resultado en ese instrumento será "SI: sin información".

[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, los 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; por otra parte, para realizar el escalamiento, el sustentante debe, al menos, haber alcanzado un acierto en el examen; en caso contrario, se reportará como cero y obtendrá N I. 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 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 instrumento el punto de corte refiera a número de aciertos diferente. En este ejemplo la distribución de las puntuaciones va de 65 a 125 puntos.



4.1.2 Con relación a los cuestionarios que integran la *Etapa 1. Informe de responsabilidades profesionales*

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos, se integrará la información y se definirán cuatro categorías que indicarán el nivel de cumplimiento del sustentante en las responsabilidades profesionales de su función[3]. Cada una de estas categorías tendrá asociada una cantidad de puntos que, como posteriormente se indicará, se adicionará a la puntuación total ponderada, considerando el siguiente orden:

NI: 0 puntos

NII: 1 punto

NIII: 2 puntos

NIV: 3 puntos

Cada cuestionario contribuirá con el 50% de la puntuación de la etapa 1, de tal forma que, en caso de faltar las respuestas de alguno de los dos cuestionarios, la puntuación de la etapa será igual a la puntuación que aporta el cuestionario del que se cuente con información.

En ningún caso, por sí mismo, la omisión de alguno de los dos cuestionarios que considera esta etapa de la evaluación **será causal de un resultado Insuficiente**. Lo anterior porque se trata de reconocer y estimular la participación genuina de los sustentantes y autoridades superiores.

4.2 Resultado global y procedimiento para la conformación de los grupos de desempeño

4.2.1 El resultado global

Para determinar el resultado global de la calificación de los sustentantes, deberán integrarse los resultados de los instrumentos considerados en las tres etapas que conforman el diseño de la evaluación, conforme a los siguientes criterios:

1) Sustentar los tres instrumentos que constituyen las etapas 2 y 3

2) Obtener al menos NII en por lo menos dos de los tres instrumentos de las etapas 2 y 3

[3] Para mayores detalles sobre el procedimiento para el escalamiento de las puntuaciones de los cuestionarios, la integración de la información y la asignación de niveles de cumplimiento en la etapa 1, consultar el anexo técnico.

Cuando no se cumpla con los criterios 1 y 2, no aplicarán los numerales 3, 4 y 5

3) Una vez que se verifica el cumplimiento de los criterios 1 y 2, se calcula la puntuación total ponderada del sustentante, es decir, se pondera[4] el resultado obtenido en los tres instrumentos de las etapas 2 y 3 bajo el siguiente esquema:

a. Etapa 2. Proyecto de enseñanza, 60%

b. Etapa 3. Examen de conocimientos y habilidades didácticas, 40%

- Para el caso de docentes:
  - Examen de conocimientos disciplinares, 20%
  - Examen de habilidades didácticas, 20%
- Para el caso de técnicos docentes:
  - Examen de conocimientos científicos y tecnológicos, 20%
  - Examen de habilidades didácticas, 20%

4) Se adiciona el resultado obtenido en la etapa 1, de acuerdo con el nivel de cumplimiento alcanzado: NI (0 puntos), NII (1 punto), NIII (2 puntos), o bien NIV (3 puntos).

5) Se asigna el resultado global de la evaluación, que integra los resultados parciales de todo el proceso.

4.2.2 La conformación de los grupos de desempeño

El resultado "Suficiente"

Para alcanzar al menos un resultado suficiente en la evaluación, se deben cumplir los siguientes criterios:

- Sustentar los tres instrumentos que constituyen las etapas 2 y 3
- Obtener al menos NII en por lo menos dos de los tres instrumentos de las etapas 2 y 3
- Obtener al menos 100 puntos en la escala de calificación global

Los **grupos de desempeño** estarán conformados únicamente por los sustentantes que obtengan, al menos, un resultado "Suficiente" en la evaluación:

Criterios para formar parte de un grupo de desempeño	
Grupo de desempeño	Puntuación en escala de calificación global
Suficiente	Al menos 100[5] puntos
Bueno	Al menos PC_2G puntos
Destacado	Al menos PC_3G puntos
Excelente	Al menos PC_4G puntos

El resultado "Insuficiente"

En los siguientes casos se asignará el resultado "Insuficiente", y por lo tanto el docente o técnico docente **no formará parte de los grupos de desempeño**, **pero recibirá la retroalimentación que corresponda**:

- No sustente los tres instrumentos que constituyen las etapas 2 y 3.
- **No obtenga** al menos NII en por lo menos dos de los tres instrumentos que constituyen las etapas 2 y 3.
- No obtenga **al menos** 100 puntos en la escala de calificación global.

En los dos primeros casos no se dará puntuación global al docente.

En los tres casos los sustentantes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan presentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

---

[4] Se traduce como la cantidad de puntos en escala INEE multiplicada por 0.60, 0.20 y 0.20, respectivamente. La puntuación de la Etapa 3 se calcula considerando que cada uno de los dos exámenes que la componen aporta el 50%. Para mayores detalles sobre el algoritmo para el cálculo de la puntuación global, consultar el anexo técnico.

[5] PC_1G siempre es igual a 100, toda vez que el primer punto de corte en los instrumentos considerados en las etapas 2 y 3 siempre es 100. Para mayores detalles sobre el algoritmo para el cálculo de los puntos de corte en la escala de calificación global, consultar el anexo técnico.

El resultado "No se presentó a la evaluación"

Para el caso en que el sustentante no presente ninguno de los instrumentos de las etapas 2 y 3 considerados en el diseño de la evaluación, ni el cuestionario de autoevaluación de la etapa 1, en el resultado de la evaluación se indicará: "No se presentó a la evaluación" y en cada instrumento sólo se le asignará "NP: No presentó". Asimismo, debido a que no se cuenta con información, tampoco tendrá devolución de los instrumentos, aun cuando su autoridad inmediata haya respondido el cuestionario que le corresponde de la etapa 1.

Sobre los resultados de la evaluación

El resultado de la evaluación, tanto para los resultados "Insuficientes", como de aquellos que forman parte de un grupo de desempeño ("Suficiente", "Bueno", "Destacado" o "Excelente"), aportará información relevante para diseñar programas y acciones de capacitación, formación y acompañamiento.

Segunda sección.

Evaluación del desempeño en su segunda o tercera oportunidad del personal docente y técnico docente en Educación Media Superior

De conformidad con la Ley General del Servicio Profesional Docente, esta evaluación del desempeño en su segunda o tercera oportunidad es obligatoria y deberá llevarse a cabo en un plazo no mayor de doce meses después de haberse presentado la primera o segunda evaluación, respectivamente.

Serán sujetos a una segunda o tercera oportunidad de evaluación del desempeño exclusivamente los docentes y técnicos docentes que obtuvieron resultado insuficiente en su primera o segunda evaluación del desempeño, respectivamente.

La calificación global se estimará siguiendo el mismo modelo de calificación desarrollado en los presentes criterios técnicos (véase la primera sección). Se considerarán los resultados obtenidos en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) con base en las siguientes equivalencias:

Equivalencias para la etapa 2

Se recuperará la información de los resultados que el sustentante haya obtenido en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) en los siguientes instrumentos de evaluación:

- *Planeación didáctica argumentada*
- *Expediente de evidencias de enseñanza*

Las reglas de equivalencias serán las siguientes:

Resultado obtenido en Planeación didáctica argumentada	***Resultado obtenido en Expediente de evidencias de enseñanza***	**Resultado asignado para la etapa 2**
NII, NIII o NIV	*NII, NIII o NIV*	El nivel de desempeño más alto que haya alcanzado en cualquiera de los dos instrumentos
En cualquier resultado cuya combinación de los dos instrumentos sea: NP o NI con NP, NI, NII, NIII o NIV		Debe presentar el Proyecto de enseñanza

Equivalencias para la etapa 3

Se recuperará la información de los resultados que el sustentante haya obtenido en su anterior evaluación del desempeño (primera o segunda oportunidad, según corresponda) en los siguientes instrumentos de evaluación:

- *Examen de conocimientos disciplinares (sólo aplicó para docentes)*
- *Examen de competencias didácticas*

Las reglas de equivalencias serán las siguientes:

Resultado obtenido en Examen de conocimientos disciplinares	**Resultado asignado al Examen de conocimientos disciplinares de la etapa 3**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NP o NI	Debe presentar el Examen de conocimientos disciplinares

Resultado obtenido en Examen de competencias didácticas	**Resultado asignado al Examen de habilidades didácticas de la etapa 3**
NII, NIII o NIV	El nivel de desempeño alcanzado en el instrumento
NP o NI	Debe presentar el Examen de habilidades didácticas

Para los **técnicos docentes sujetos a su segunda oportunidad de la evaluación del desempeño, la puntuación en la etapa 3 estará dada considerando únicamente su resultado en el Examen de habilidades didácticas**, por lo tanto, su resultado global se determinará integrando los resultados de los instrumentos considerados en las tres etapas que conforman el diseño de la evaluación, conforme a los siguientes criterios:

1) Sustentar los dos instrumentos que constituyen las etapas 2 y 3

2) Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3

Cuando no se cumpla con los criterios 1 y 2, no aplicarán los numerales 3, 4 y 5

3) Una vez que se verifica el cumplimiento de los criterios 1 y 2, se calcula la puntuación total ponderada del sustentante, es decir, se pondera[6] el resultado obtenido en los dos instrumentos de las etapas 2 y 3 bajo el siguiente esquema:

a. Etapa 2. Proyecto de enseñanza, 60%

b. Etapa 3. Examen de habilidades didácticas, 40%

4) Se adiciona el resultado obtenido en la etapa 1, de acuerdo con el nivel de cumplimiento alcanzado: NI (0 puntos), NII (1 punto), NIII (2 puntos), o bien NIV (3 puntos).

5) Se asigna el resultado global de la evaluación, que integra los resultados parciales de todo el proceso.

De esta forma, para alcanzar al menos un **resultado Suficiente** en la evaluación, estos sustentantes deben cumplir los siguientes criterios:

- Sustentar los dos instrumentos que constituyen las etapas 2 y 3
- Obtener al menos NII en por lo menos uno de los dos instrumentos de las etapas 2 y 3
- Obtener al menos 100 puntos en la escala de calificación global

Asimismo, en los siguientes casos se asignará el resultado **Insuficiente** y, por lo tanto, el sustentante **no formará parte de los grupos de desempeño**, **pero recibirá la retroalimentación que corresponda**:

- No sustente los dos instrumentos que constituyen las etapas 2 y 3.
- **No obtenga** al menos NII en por lo menos uno de los dos instrumentos que constituyen las etapas 2 y 3.
- No obtenga **al menos** 100 puntos en la escala de calificación global.

En los dos primeros casos no se dará puntuación global al sustentante.

En los tres casos los sustentantes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan presentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

Finalmente, cualquier situación no prevista en los presentes criterios técnicos será analizada por la Junta de Gobierno 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* y toda vez que los cuestionarios que constituyen la etapa 1 de este proceso tienen como finalidad recabar información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función profesional, y **únicamente** pueden ser considerados para **adicionar puntos al sustentante en su calificación global, la cual está en función de los resultados alcanzados en los instrumentos que constituyen las etapas 2 y 3**, es fundamental señalar que, en ningún caso, **se puede considerar solamente un instrumento** para integrar la calificación de los sustentantes conforme al diseño de la evaluación, es decir:

[6] Se traduce como la cantidad de puntos en escala INEE multiplicada por 0.60 y 0.40, respectivamente. Para mayores detalles sobre el algoritmo para el cálculo de la puntuación global, consultar el anexo técnico.

Ninguna decisión que tenga consecuencias importantes sobre los individuos o instituciones, se basará únicamente en los resultados de sólo un instrumento de evaluación, por lo cual, deberán considerarse otras fuentes confiables de información que incrementen la validez de las decisiones que se tomen.

Lo anterior debido a que la evidencia empírica que resulte del análisis psicométrico de los instrumentos de la segunda y tercera etapa de la evaluación del desempeño del personal docente y técnico docente debe mostrar que, una vez que éstos fueron aplicados, cumplen con los criterios técnicos establecidos por el Instituto, de esta forma 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 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 ni el centro de trabajo 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 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 calificación que se asigna 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 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 del sustentante en cada aspecto a evaluar para emitir su calificación definitiva 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 del punto 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 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:

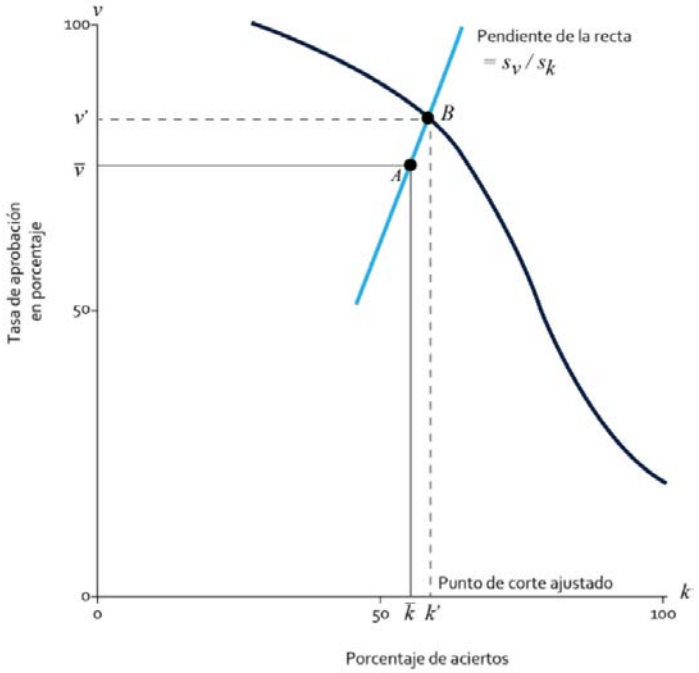
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas ($\bar{k}$, $\bar{v}$), (ver siguiente figura).

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: $v = \left(^{s_v}/_{s_k}\right)(k - \bar{k}) + \bar{v}$.

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 de los instrumentos considerados en las etapas 2 y 3

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 II) 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.

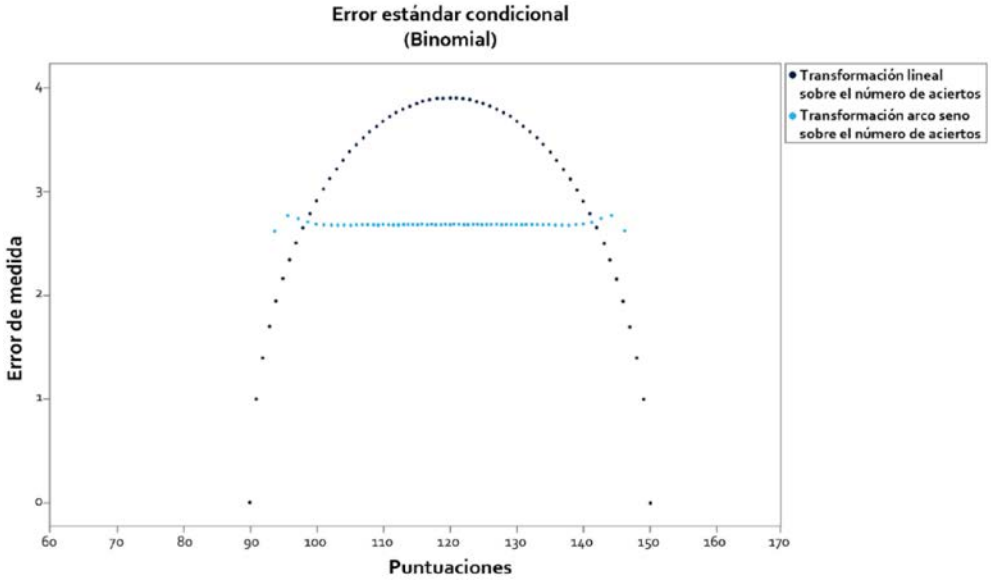
Finalmente, es importante destacar que para que se lleve a cabo el escalamiento, el sustentante debió alcanzar, al menos, un acierto en el instrumento de evaluación en cuestión. De no ser así, se reportará como cero y el resultado será N I.

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

Como se comentó, 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 II (N II) en los instrumentos de evaluación, se ubicará en el valor 100. Para determinarla se empleará la siguiente ecuación:

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

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 II.

El valor de Q dependerá 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 (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 puntuaciones 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 1 (P_{A1i}), 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 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, 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 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, el punto 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.

$\boldsymbol{\alpha}$ 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

$\sigma_X^2 =$ 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:

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

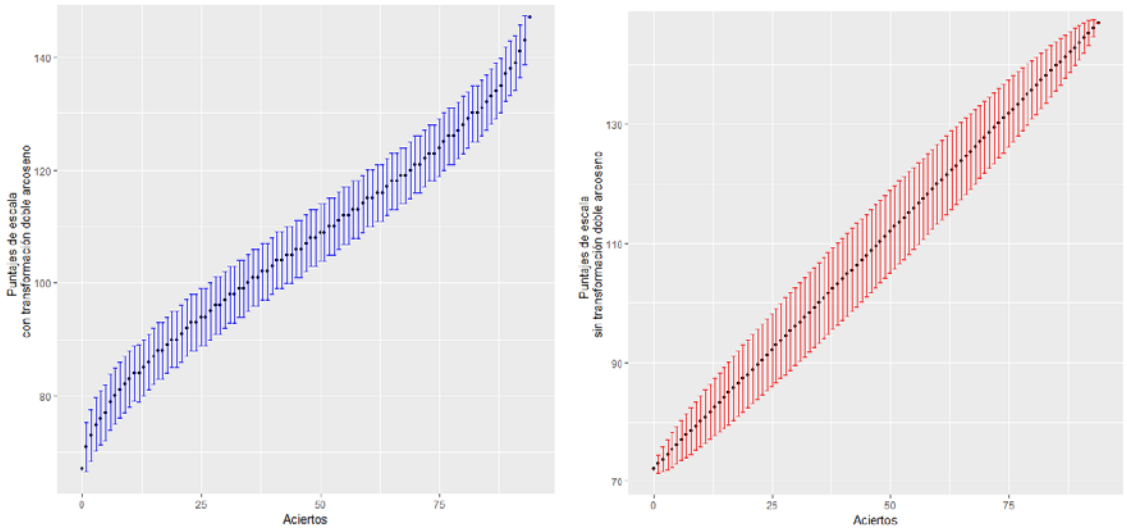
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

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.

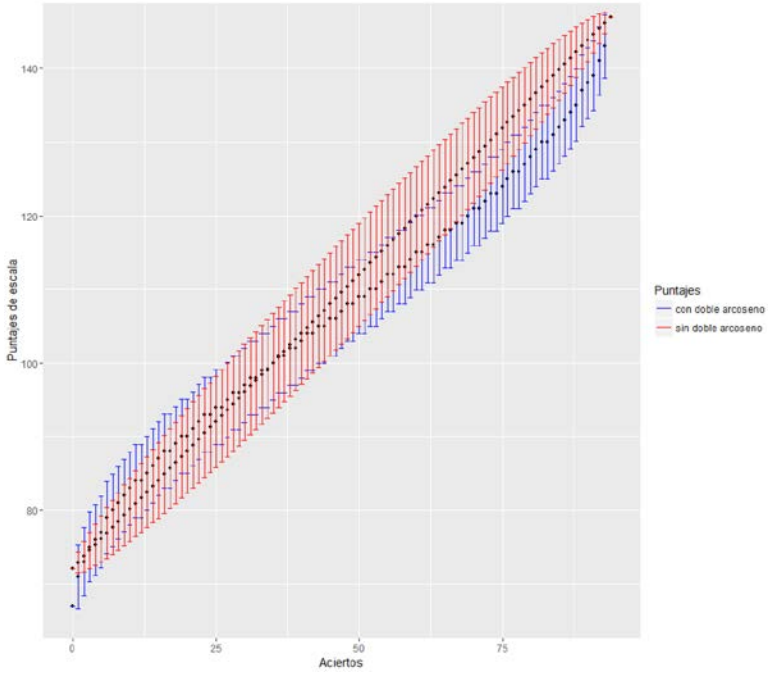
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 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 II (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

Como ya se indicó en el cuerpo del documento, 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

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

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

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

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

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

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

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

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

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

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.

Es importante señalar que para los puntajes que se les aplique la equiparación $x_e = b_1x + b_0$, 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).

Procedimiento para el escalamiento de las puntuaciones de los cuestionarios de la etapa 1

La etapa 1 de este proceso de evaluación está constituida por dos cuestionarios, cuya función es obtener información sobre el nivel de cumplimiento de las responsabilidades profesionales asociadas a la función:

a) Cuestionario respondido por el sustentante.

b) Cuestionario respondido por su autoridad inmediata.

Con base en las respuestas que el sustentante y su autoridad inmediata den a los cuestionarios, se realizará el escalamiento de las puntuaciones para cada uno de ellos.

La escala de puntuaciones de cada cuestionario se ubicará en el intervalo [0, 50], si un cuestionario no es presentado se le asignará una puntuación de cero. Ambos cuestionarios serán escalados utilizando el modelo de crédito parcial. Para que el rango de puntuaciones vaya de 0 a 50, las puntuaciones que se obtengan con el modelo se escalarán linealmente y se redondearán al entero más próximo, utilizando el criterio de que puntuaciones con cinco décimas o más, suben al siguiente entero.

De esta forma, la puntuación alcanzada en la etapa 1 será calculada como la suma de las puntuaciones de ambos cuestionarios, por lo que se ubicará en el intervalo [0, 100].

La asignación del nivel de cumplimiento en la etapa 1 y la cantidad de puntos que se adicionan a la puntuación total del sustentante, será con base en la siguiente tabla:

Suma de las puntuaciones de ambos cuestionarios	Nivel de cumplimiento	Puntos que se adicionan
De 0 a 25	NI	0
De 26 a 50	NII	1
De 51 a 75	NIII	2
De 76 a 100	NIV	3

Algoritmo para el cálculo de la puntuación global

Una vez que se ha verificado que el sustentante presentó los tres instrumentos que constituyen las etapas 2 y 3 del proceso de evaluación y que obtuvo al menos NII en por lo menos dos de ellos, se procede a calcular la puntuación global con base en el siguiente esquema:

Etapa 2. Proyecto de enseñanza, 60%

Etapa 3. Examen de conocimientos y habilidades didácticas, 40%

- Para el caso de docentes:
  - Examen de conocimientos disciplinares, 20%
  - Examen de habilidades didácticas, 20%

- Para el caso de técnicos docentes[7]:
  - Examen de conocimientos científicos y tecnológicos, 20%
  - Examen de habilidades didácticas, 20%

$$\boldsymbol{G_i = 0.60 * P_{1i} + 0.20 * P_{2i} + 0.20 * P_{3i} + P_{Ei}}$$

$G_i = Puntuación\ global\ que\ alcanza\ el\ sustentante\ i\ en\ la\ evaluación$

$P_{1i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Proyecto\ de\ enseñanza$

$P_{2i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ de\ conocimientos\ disciplinares\ (docentes)\ o\ Examen\ de\ conocimientos\ científicos\ y\ tecnológicos\ (técnicos\ docentes)$

$P_{3i} = Puntuación\ en\ escala\ INEE\ que\ alcanza\ el\ sustentante\ i\ en\ el\ instrumento\ Examen\ de\ habilidades\ didácticas$

$P_{Ei} = 0,1,2,3\ (Puntuación\ que\ se\ adiciona\ con\ base\ en\ el\ resultado\ del\ sustentante\ i\ en\ la\ etapa\ 1)$

Algoritmo para el cálculo de los puntos de corte en la escala de calificación global

Los puntos de corte 1, 2 y 3 en la escala global se calcularán considerando los puntos de corte establecidos en los instrumentos utilizados en las etapas 2 y 3, con base en el siguiente algoritmo[8]:

$$\boldsymbol{PC_iG = 0.60 * PC_iP + 0.20 * PC_iEC + 0.20 * PC_iEH}$$

$i = 1, 2, 3$

$PC_iG = Punto\ de\ corte\ i\ en\ la\ escala\ de\ calificación\ global$

$PC_iP = Punto\ de\ corte\ i\ establecido\ en\ el\ Proyecto\ de\ enseñanza$

$PC_iEC = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ de\ conocimientos\ disciplinares\ (docentes)\ o\ Examen\ de\ conocimientos\ científicos\ y\ tecnológicos\ (técnicos\ docentes)$

$PC_iEH = Punto\ de\ corte\ i\ establecido\ en\ el\ Examen\ de\ habilidades\ didácticas$

Finalmente, el cuarto punto de corte en la escala de calificación global (PC_4G) que permite diferenciar el nivel de desempeño "Destacado" del nivel de desempeño "Excelente" se calculará con base en lo siguiente:

$$\boldsymbol{PC_4G = \frac{1}{3}(PC_3G + [2 * max\{PINEE\}])}$$

$max\{PINEE\} = Puntuación\ global\ máxima\ posible\ en\ escala\ INEE$

$PC_3G = Punto\ de\ corte\ 3\ en\ la\ escala\ de\ calificación\ global\ (previamente\ calculado)$

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.

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.

Bentler, P. M. (2006). EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

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.

Brennan, R. L. (2012). Scaling PARCC Assessments: Some considerations and a synthetic data example en: http://parcconline.org/about/leadership/12-technical-advisory-committee

[7] Para el caso de los técnicos docentes sujetos a su segunda oportunidad de la evaluación del desempeño, la etapa 3 considera únicamente el Examen de habilidades didácticas, por lo que la ponderación asociada a la puntuación del sustentante en dicho examen es 40% y la asociada al examen de conocimientos científicos y tecnológicos es 0%.

[8] Para el caso de los técnicos docentes sujetos a su segunda oportunidad de la evaluación del desempeño, la etapa 3 considera únicamente el Examen de habilidades didácticas, por lo que la ponderación del PC_iEH es 40% y del PC_iEC es cero.

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

Downing, SM (2004). Reliability: On the reproducibility of assessment data. Med Educ; 38(9):1006-1012. 21

Holland, P. W., & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), Statistical models for test equating, scaling, and linking (pp. 89–107). New York, NY: Springer

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. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Masters, Geoff (1982). A Rasch model for Partial Credit Scoring. Psychometrika-vol. 47, No. 2.

Muñiz, José (2003): Teoría clásica de los test. Ediciones pirámide, Madrid.

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

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

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

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

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

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

Shun-Wen Chang (2006) Methods in Scaling the Basic Competence Test, *Educational and Psychological Measurement,* 66 (6) 907-927

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

Stemler, E. & Tsai, J. (2008). *Best Practices in Interrater Reliability Three Common Approaches* in Best practices in quantitative methods (pp. 29–49). SAGE Publications, Inc.

Thompson, Bruce ed. (2003): Score reliability. Contemporary thinking on reliability issues. SAGE Publications, Inc.

Wilson, Mark (2005). Constructing measures. An ítem response modeling approach. Lawrence Erlbaum Associates, Publishers.

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.

Wu, Margaret & Adams, Ray (2007). Applying the Rasch Model to Psycho-social measurement. A practical approach. Educational measurement solutions, Melbourne.

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

Ciudad de México, a veintiocho de septiembre de dos mil diecisiete.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil diecisiete, celebrada el veintiocho de septiembre de dos mil diecisiete. Acuerdo número SOJG/09-17/07,R. El Consejero Presidente, **Eduardo Backhoff Escudero**.- Rúbrica.- Los Consejeros: **Teresa Bracho González**, **Gilberto Ramón Guevara Niebla, Sylvia Irene Schmelkes del Valle, Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 457558)