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 de la evaluación del desempeño para personal a cargos con funciones de Dirección en la Educación Media Superior, ciclo escolar 2018-2019.

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 DE LA EVALUACIÓN DEL DESEMPEÑO PARA PERSONAL A CARGOS CON FUNCIONES DE DIRECCIÓN EN LA EDUCACIÓN MEDIA SUPERIOR, CICLO ESCOLAR 2018-2019

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).

Así, y con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 3, 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 cuarto grupo de docentes y técnicos docentes y del personal que presenta su segunda y tercera oportunidad, así como del personal con funciones de Dirección en Educación Media Superior en el ciclo escolar 2018-2019. LINEE-08-2018, publicados el 7 de mayo del 2018 donde se establecen las directrices para llevar a cabo la evaluación del desempeño del personal con funciones de dirección en el ciclo escolar 2018-2019 en Educación Media Superior, 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 de la evaluación del desempeño a cargos con funciones de Direccion en la Educación Media Superior, ciclo escolar 2018-2019.

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 en cuatro apartados: 1) Características generales de los instrumentos para evaluar el desempeño del personal con funciones de dirección; 2) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 3) Procedimiento para el establecimiento del punto de corte y estándar de desempeño de los instrumentos de evaluación considerados en las etapas 2 y 3; 4) Resultado de la evaluación del desempeño por: instrumento, etapa y resultado global. En la parte final se presenta un Anexo técnico con información detallada de algunos aspectos 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. **Formas de un instrumento**: Dos o más versiones de un instrumento que se consideran equivalentes, pues se construyen con los mismos contenidos y especificaciones estadísticas.

XXII. **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.

XXIII. **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.

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

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

XXVI. **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.

XXVII. **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.

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

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

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

XXXI. **Porcentaje de acuerdos inter-jueces:** Medida del grado en que dos jueces coinciden en la puntuación asignada a un sustentante cuyo desempeño es evaluado a través de una rúbrica.

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

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

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

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

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

XXXVII. **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.

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

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

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

La evaluación del desempeño es un proceso integrado que incluye varios instrumentos que dan cuenta de los diferentes aspectos descritos 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

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 intervención de la gestión del director

El proyecto de intervención de la gestión del personal con funciones de dirección es un instrumento que permite evaluar el desempeño o labor de quienes han llevado a cabo funciones de dirección en la Educación Media Superior. Consiste en la elaboración de un diagnóstico y un proyecto de intervención y la reflexión sobre su práctica. El proyecto de intervención de la gestión del director se constituye en tres momentos:

Momento 1. Diagnóstico y proyecto de intervención

Momento 2. Intervención

Momento 3. Reflexión sobre su práctica

Etapa 3. Examen de conocimientos curriculares y de normatividad para el director

Este instrumento evalúa los conocimientos, capacidades y habilidades de este personal para afrontar y resolver diversas situaciones de la práctica profesional propias de la función directiva. El propósito de la evaluación es: evaluar el nivel de dominio que posee el director de los conocimientos normativos y de la gestión, en términos de la función en concurso, para el mejor funcionamiento del plantel.

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 en este documento, 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 corregida 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 dos niveles de desempeño posibles: Nivel I (NI) y Nivel II (NII). 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 intervención de la gestión del director

Nivel de desempeño	Descriptor
Nivel I (N I)	En este nivel, el personal con funciones de dirección, como máximo, es capaz de elaborar un proyecto de intervención de gestión, pero carece de un orden lógico entre sus elementos. Además, algunos elementos de la estrategia de trabajo, la organización escolar y las acciones de seguimiento están desvinculadas de la prioridad educativa seleccionada. Con respecto al desarrollo del plan de trabajo, es capaz de describir acciones como las de atención a la prioridad educativa, para los ámbitos de gestión escolar y de organización, y para la generación de ambientes favorables para el aprendizaje, la sana convivencia, la inclusión, la equidad y la diversidad, las cuales no siempre son congruentes con algunos elementos del proyecto de intervención de gestión o con la prioridad educativa seleccionada. En lo que corresponde al análisis y reflexión de su gestión directiva, puede referir algunos resultados de su intervención y relacionarlos con algunos logros de los objetivos y metas planteadas para la atención de la prioridad educativa. Asimismo, puede mencionar acciones de mejora de su práctica basadas en su experiencia o en sus conocimientos; mediante la descripción de algunas fortalezas y aspectos a mejorar que no siempre toman en cuenta los resultados de su intervención.
Nivel II (N II)	En este nivel, además de lo implicado en el nivel I, el personal con funciones de dirección es capaz de justificar cómo la estrategia de trabajo establecida en el proyecto es congruente para atender a la prioridad educativa seleccionada, considerando las características específicas que describió en el diagnóstico escolar y los resultados de las evaluaciones internas y externas. Con respecto al desarrollo del proyecto de intervención, es capaz de explicar, de manera lógica y articulada, la forma en la que durante la implementación de su proyecto retomó los resultados de las evaluaciones internas y externas para vincularlos con la estrategia de trabajo establecida y para la atención de la prioridad educativa seleccionada. Asimismo, puede explicar cómo es que las acciones de la estrategia de trabajo implementadas permiten establecer situaciones escolares para la generación de ambientes favorables para el aprendizaje, la sana convivencia y la inclusión. En lo que corresponde al análisis y reflexión de su gestión directiva, es capaz de describir, de manera detallada, las acciones de mejora que implementó durante su intervención, basadas en su experiencia y en sus conocimientos. Asimismo, puede explicar la relación entre los resultados de su intervención y los logros de los objetivos y metas planteadas para la atención de la prioridad educativa; y puede justificar sus fortalezas y las acciones concretas para mejorar los resultados de su intervención, congruentes con el contexto en que se desempeña.

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

Nivel de desempeño	Descriptor
Nivel I (N I)	En este nivel, el personal con funciones de dirección, como máximo, es capaz de identificar planes y programas de estudio vigentes, reconocer los enfoques y vincularlos con actividades de clase de los docentes para promover la atención a la diversidad. Sin embargo, desconoce aspectos básicos del Padrón de Calidad del Sistema Nacional de Educación Media Superior, relacionados con las etapas, requisitos y procedimientos para ingresar y mantener la permanencia en el Programa. Asimismo, muestra dificultades para reconocer estrategias encaminadas a fomentar los principios éticos de responsabilidad y corresponsabilidad entre los miembros de la comunidad escolar. Muestra dificultad para identificar acciones directivas de retroalimentación académica a los docentes del plantel, a partir de los resultados que éstos obtuvieron en alguna de las etapas del proceso de evaluación del desempeño del Servicio Profesional Docente, y las estrategias orientadas a su formación continua; así como para reconocer estrategias de capacitación orientadas a la mejora de las prácticas docente y administrativa. Manifiesta problemas para reconocer la estrategia directiva que promueve el desarrollo integral de los alumnos, la participación de toda la comunidad escolar en la construcción de ambientes positivos para el aprendizaje, el fomento de estilos de vida saludable y el desarrollo humano de los estudiantes y un ambiente escolar de respeto y tolerancia a la diversidad. Asimismo, no logra identifica las acciones directivas para solucionar conflictos entre los diferentes actores de la comunidad escolar.
Nivel II (N II)	En este nivel, además de lo implicado en el nivel I, el personal con funciones de dirección es capaz de identificar estrategias de intervención que, desde el ámbito de su función y considerando las necesidades de aprendizaje y los resultados de las evaluaciones internas y externas de la escuela, favorecen la mejora de los aprendizajes de los alumnos. Asimismo, reconoce la acción directiva que da solución a problemas que implican la administración de recursos humanos, materiales y financieros, con apego a la normatividad vigente de operación escolar y los protocolos de actuación para la atención a alumnos con necesidades educativas especiales o en situación de riesgo y vulnerabilidad; además, identifica mecanismos de atención y comunicación con los padres de familia. Distingue acciones directivas de seguimiento a las metas establecidas en el Plan de Mejora Continua, en relación con: el abandono escolar, la eficiencia terminal, el aprovechamiento escolar y la normalidad mínima de operación escolar, considerando las características del plantel y del entorno. Asimismo, es capaz de identificar la estrategia directiva que promueve el desarrollo integral de los alumnos, la participación de toda la comunidad escolar en la construcción de ambientes positivos para el aprendizaje, el fomento de estilos de vida saludable y el desarrollo humano de los estudiantes y un ambiente.

Segunda fase

En esta fase se establece el punto 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 el nivel de desempeño NII. 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 por: instrumento, etapa 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 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.

Escala utilizada para reportar los resultados

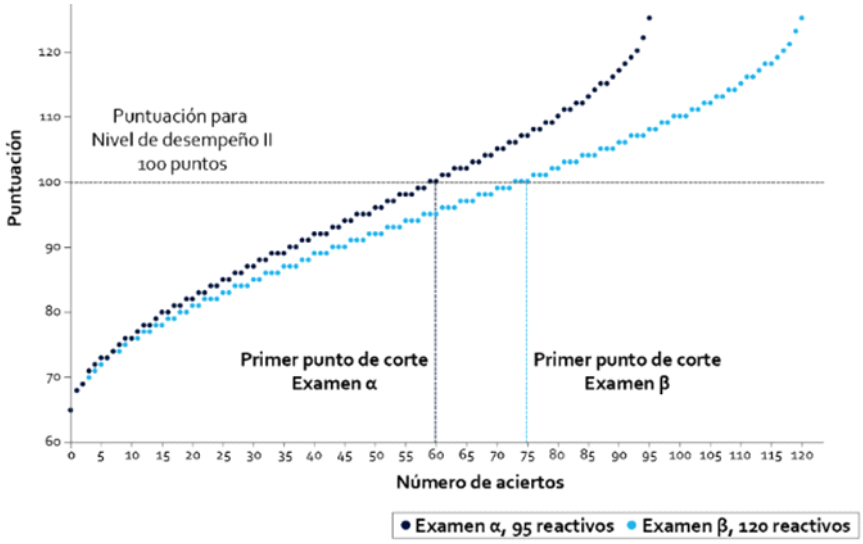
En cada plan de evaluación es indispensable definir la escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. En 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[1]).

[1] Pueden encontrarse ligeras variaciones en este rango debido a que la escala es aplicable a múltiples instrumentos con características diversas, tales como: longitudes de la prueba, nivel de precisión, tipo de instrumento, entre otras. Por otra parte, para realizar el escalamiento, el sustentante debe haber presentado el instrumento de evaluación, y al menos, haber obtenido un acierto en el examen; en caso de no haber obtenido por lo menos un acierto, se reportará como cero y obtendrá NI. Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

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

En la siguiente gráfica puede observarse el número de aciertos obtenido en dos instrumentos de longitudes diferentes y con puntos de corte distintos que, a partir del escalamiento, es posible graficar en una misma escala, trasladando el 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 de autoevaluación, respondido por el personal con funciones de dirección

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[2]. 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:

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**

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.

[2] 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.

La omisión de alguno de los dos cuestionarios que considera esta etapa de la evaluación **no será causal de un resultado Insuficiente**. Lo anterior porque se trata de reconocer y estimular la participación genuina de los sustentantes y autoridades superiores.

Consultar excepción en la sección "*El resultado No se presentó a la evaluación**".***

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

Para el cálculo del resultado global y la formación de los grupos de desempeño se tomarán en cuenta los siguientes instrumentos:

- Etapa 1: *Informe de responsabilidades profesionales*
  - Cuestionario de autoevaluación
  - Cuestionario respondido por su autoridad inmediata
- Etapa 2: *Proyecto de intervención de la gestión del director*
- Etapa 3: *Examen de conocimientos curriculares y de normatividad para el director*

En el caso en que el sustentante no presente alguno de los instrumentos de evaluación que le corresponda de las etapas 1, 2 o 3 su resultado en ese instrumento será "NP: no se presentó" y únicamente tendrá la retroalimentación en aquellos instrumentos en los que haya participado, y de los que se cuente con información. 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 "NP: no se presentó".

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[3] el resultado obtenido en los dos instrumentos de las etapas 2 y 3 bajo el siguiente esquema:
   a. Etapa 2. Proyecto de intervención de la gestión del director, 60%
   b. Etapa 3. Examen de conocimientos curriculares y de normatividad para el director, 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) El resultado global de la evaluación se asigna al integrar los resultados parciales de todo el proceso, de tal forma que los criterios para los posibles resultados de la evaluación son los siguientes:

Resultado de la evaluación	Criterios
Cumple con la función de dirección	● 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.
Insuficiente	● No sustente al menos un instrumento de los 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.
No se presentó a la evaluación	En caso de que el sustentante no presente NINGUNO de los instrumentos considerados en las etapas 2 y 3, ni el cuestionario de autoevaluación de la etapa 1

[3] Para mayores detalles sobre el cálculo de la puntuación global, consultar el anexo técnico.

En los casos donde el resultado de la evaluación sea ***No se presentó a la evaluación o*** en aquellos casos donde el sustentante no presente alguno de los instrumentos de la etapa 2 o 3, o no obtenga al menos un nivel NII en uno de los dos instrumentos de las etapas 2 y 3, 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, pero no se calculará una puntuación global.

El resultado No se presentó a la evaluación

En los casos donde el resultado de la evaluación sea ***No se presentó a la evaluación***, en cada instrumento sólo se asignará "NP: no se presentó", asimismo, debido a la falta de información para el cálculo de puntajes, tampoco existirá retroalimentación de los instrumentos que constituyen el proceso de evaluación del desempeño; aun cuando la autoridad inmediata haya respondido el cuestionario de la etapa 1, en cuyo caso se le asignará "NP: no se presentó" y el resultado de la etapa será equivalente a "NP: no se presentó".

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

Consideración final

Es importante destacar que el resultado global de la evaluación es el que debe considerarse como el marco de interpretación para el cumplimiento de la función de dirección, ya que integra los resultados obtenidos en los instrumentos que constituyen las etapas del proceso de evaluación.

Consulta pública de resultados

Al igual que los informes individuales de resultados, la consulta pública de los mismos se realizará en la página de la CNSPD y, para estos últimos se deberá considerar su presentación ordenando los resultados de manera descendente, primero por grupo de desempeño y posteriormente la puntuación global.

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 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:

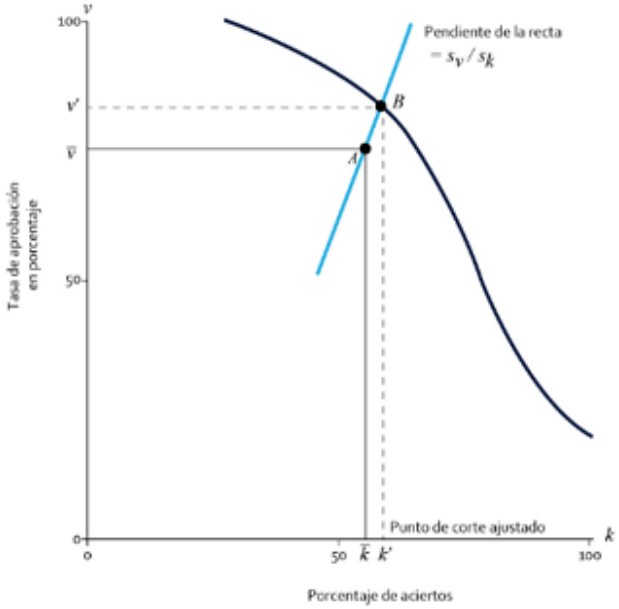
1. Se calcula el promedio de k ($\bar{k}$), y de v ($\bar{v}$). Ambos valores generan el punto A con coordenadas $(\bar{k}, \bar{v})$, (ver siguiente figura).

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

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 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 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).

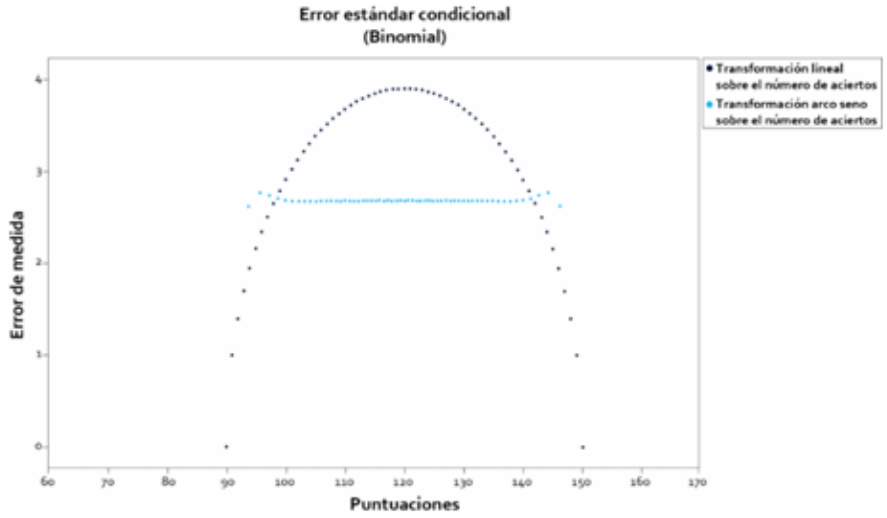
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

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*:

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

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 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:

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

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 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:

$$P_{A1i} = P_i * \frac{k_{A1i}}{k_i} \tag{3}$$

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

$$P_{A2i} = P_i - P_{A1i} \tag{4}$$

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:

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

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:

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

Donde:

X es una variable aleatoria asociada a los puntajes

n es el número de reactivos del instrumento

KR21 es el coeficiente de Kuder-Richardson.

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

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

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

σ_X^2 = varianza de las puntuaciones en el instrumento

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

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

De ahí que:

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

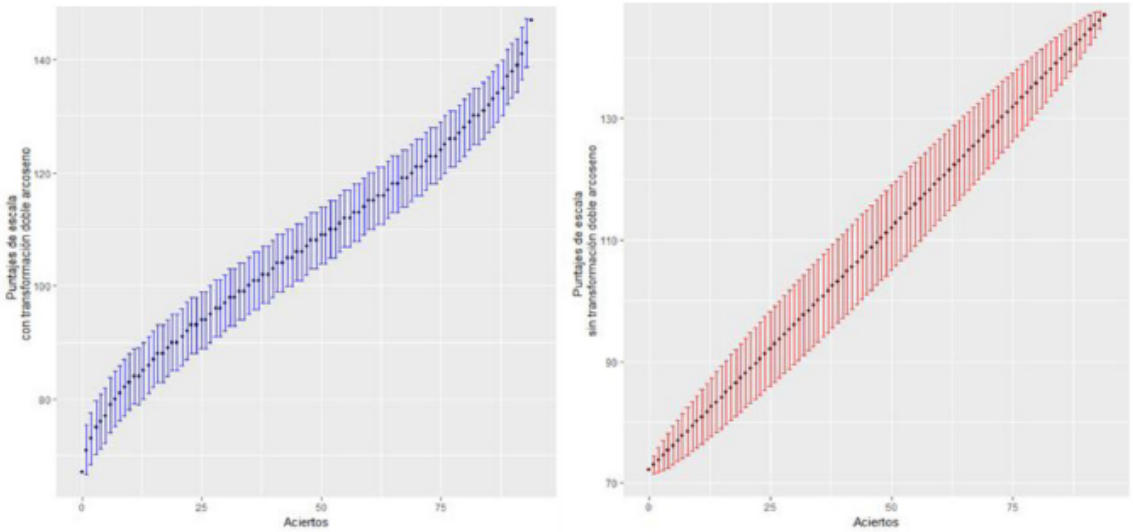
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

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

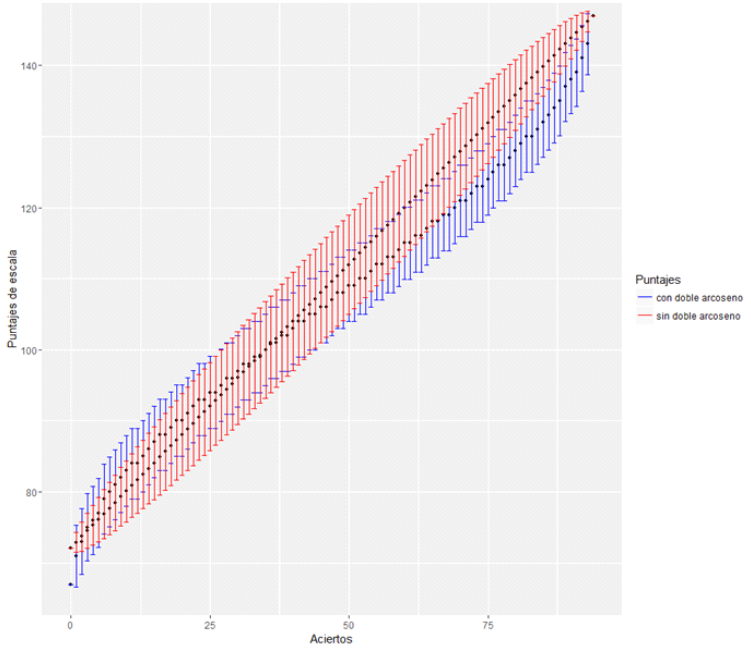
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan,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

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



Se observa que al aplicar la transformación doble arcoseno se mide con similar precisión la mayoría de los puntajes de la escala, a diferencia de cuando no se aplica dicha transformación, además de que en el punto de corte para alcanzar el nivel de desempeño 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

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

y

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

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

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

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

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

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

Donde s denota la población sintética y

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

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

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

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

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

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

y

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

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

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

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

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

Es importante señalar que para los puntajes que se les aplique la equiparación $x_e = b_1x + b_0$, 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 de autoevaluación, respondido por el personal con funciones de dirección.

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]. Para que el rango de puntuaciones vaya de 0 a 50, las puntuaciones que se obtengan se escalarán linealmente y se redondearán al entero más próximo, utilizando el criterio que 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[4], 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

Estos puntos se adicionan a la puntuación global únicamente en los casos que cumplan 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

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 intervención de la gestión del director, 60%

Etapa 3. Examen de conocimientos curriculares y de normatividad para el director, 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 intervención de la gestión del director*

P_{2i} = *Puntuación en escala INEE que alcanza el sustentante i en el instrumento Examen de conocimientos*

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

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.

[4] Cuando un cuestionario no fue respondido se le asignará “NP: no presento”.

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.

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 item 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.

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

Ciudad de México, a veintisiete de septiembre de dos mil dieciocho.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil dieciocho, celebrada el veintisiete de septiembre de dos mil dieciocho. Acuerdo número **SOJG/09-18/07,R**.- La Consejera Presidenta, **Teresa Bracho González.-** Los Consejeros: **Bernardo Hugo Naranjo Piñera, Sylvia Irene Schmelkes del Valle** y **Patricia Gabriela Vázquez del Mercado Herrera.**

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

(R.- 474123)