

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 docente y técnico docente en Educación Media Superior en el ciclo escolar 2016-2017, así como para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente en este mismo tipo educativo.

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 2016-2017, ASÍ COMO PARA LLEVAR A CABO LA EVALUACIÓN DEL DESEMPEÑO EN SU SEGUNDA OPORTUNIDAD DEL PERSONAL DOCENTE EN ESTE MISMO TIPO EDUCATIVO.

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 III, incisos b), f) y h) y VI, 52, 53 y 54 de la Ley General del Servicio Profesional Docente; 22, 28, fracción X, 38, fracciones IX y XXII de la Ley del Instituto Nacional para la Evaluación de la Educación; de los Lineamientos para llevar a cabo la evaluación del desempeño del personal docente, técnico docente y de quienes ejerzan funciones de dirección y supervisión en educación básica y media superior en el ciclo escolar 2016-2017 (LINEE-09-2016), así como de los Lineamientos para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente y técnico docente en Educación Básica y Media Superior (LINEE-10-2016), 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 2016-2017, así como para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente en este mismo tipo educativo.

Los presentes Criterios técnicos y de procedimiento tienen como finalidad establecer los referentes necesarios para garantizar la validez, confiabilidad y equidad de los resultados de los procesos de evaluación. Su contenido se organiza de la siguiente manera:

Primera sección: Sobre la evaluación del desempeño para el ciclo escolar 2016-2017

Incorpora cuatro apartados: 1) Características generales de los instrumentos para evaluar el desempeño 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; 4) Resultado de la evaluación del desempeño: resultado por instrumento y resultado global.

Segunda sección: Sobre la evaluación del desempeño de quienes será su segunda oportunidad, considerando sus resultados en el proceso de evaluación del desempeño para el ciclo escolar 2015-2016

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 del instrumento tienen consecuencias importantes para las personas o las instituciones; por ejemplo, en los procesos de admisión o certificación.
- II. **Calificación:** Proceso de asignación de una puntuación o nivel de desempeño logrado a partir de los resultados de una medición.
- III. **Confiabilidad:** Calidad de las mediciones obtenidas con un instrumento, que se caracterizan por ser consistentes y estables cuando éste se aplica en distintas ocasiones.
- IV. **Constructo:** Elaboración teórica formulada para explicar un proceso social, psicológico o educativo y cuya adecuada descripción permite que sea susceptible de ser observable o medible.

- V. **Correlación punto biserial:** Medida de consistencia que se utiliza en el análisis de reactivos, indica si hay una correlación entre el resultado de un reactivo con el resultado global del examen.
- VI. **Criterio de evaluación:** Indicador de un valor aceptable sobre el cual se puede establecer o fundamentar un juicio de valor sobre el desempeño de una persona.
- VII. **Desempeño:** Resultado obtenido por el sustentante en un proceso de evaluación o en un instrumento de evaluación educativa.
- VIII. **Dificultad de un reactivo:** Indica la proporción de personas que responden correctamente el reactivo de un examen.
- IX. **Distractores:** Opciones de respuesta incorrectas del reactivo de opción múltiple, que probablemente serán elegidas por los sujetos con menor dominio en lo que se evalúa.
- X. **Dominio:** Conjunto de conocimientos, habilidades, destrezas, actitudes u otros atributos que tienen las siguientes propiedades: límites, extensión y definición. También se puede aplicar a contenidos, procedimientos u objetos.
- XI. **Educación básica:** Tipo de educación que comprende los niveles de preescolar, primaria y secundaria en todas sus modalidades, incluyendo la educación indígena, la especial y la que se imparte en los centros de educación básica para adultos.
- XII. **Educación media superior:** Tipo de educación que comprende el nivel de bachillerato, los demás niveles equivalentes a éste, así como la educación profesional que no requiere bachillerato o sus equivalentes.
- XIII. **Equiparación:** Proceso estadístico que se utiliza para ajustar las puntuaciones de las formas de un mismo instrumento, permite que las puntuaciones de una forma a otra sean utilizadas de manera intercambiable. La equiparación ajusta, por dificultad, las distintas formas que fueron construidas con contenidos y dificultad similar.
- XIV. **Error estándar de medida:** Desviación estándar de una distribución hipotética de errores de medida de una población.
- XV. **Escala:** Procedimiento para asignar números, puntuaciones o medidas 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 la calificación que obtienen los sustentantes en uno o varios instrumentos de evaluación, colocando las puntuaciones de los distintos instrumentos o formas a una escala común.
- XVII. **Especificaciones de tareas evaluativas o reactivos:** Descripción detallada de las características relevantes que se espera hagan los sujetos al sustentar el instrumento de evaluación y que es posible observar a través de las tareas evaluativas o los reactivos. Tienen el papel de guiar a los comités académicos en la elaboración y validación de las tareas evaluativas o de los reactivos y que éstos cuenten con los elementos necesarios para construirlos alineados al objeto de medida o constructo que se desea evaluar a través del instrumento.
- XVIII. **Estándar:** Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación y que son acordados por expertos en evaluación.
- XIX. **Evaluación:** Acción de emitir juicios de valor sobre un objeto, sujeto o evento que resultan de comparar los resultados de una medición u observación con un referente previamente establecido.
- 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:** Procedimiento de recolección de datos que suelen tener distintos formatos, atendiendo a la naturaleza de la evaluación, por ejemplo, instrumentos de selección de respuesta, instrumentos de respuesta construida, cuestionarios, observaciones, portafolios, entre otros.
- XXII. **Instrumento de evaluación referido a un criterio:** Instrumento que permite comparar el desempeño de las personas evaluadas, con un estándar preestablecido.
- XXIII. **Jueceo:** Método en el cual se utiliza la opinión de expertos (denominados jueces) para determinar, entre otras cosas, la pertinencia de la validez de las tareas evaluativas o de los reactivos respecto a un dominio; el establecimiento de estándares de desempeño y puntos de corte; así como la calificación de reactivos de respuesta construida.

- 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 que sus propiedades puedan 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 un instrumento de evaluación, y que refiere a lo que la persona evaluada es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.
- 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 a 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 sujeto 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 sujetos que evalúa, coincide en la puntuación asignada al desempeño de los sujetos, 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 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.
- XXXIV. Puntuación:** Número de aciertos obtenidos en un instrumento de evaluación.
- XXXV. Reactivo:** Unidad básica de medida de un instrumento de evaluación que consiste en una pregunta o instrucción que requiere una respuesta del sujeto.
- 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 de un instrumento de evaluación de respuesta construida y que consiste en la ejecución de una actividad que es susceptible de 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.

Primera sección.

Evaluación del desempeño del personal docente y técnico docente en Educación Media Superior, 2016-2017

1. Características generales de los instrumentos para evaluar el desempeño 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 ellos.

Informe de cumplimiento de responsabilidades profesionales

Este instrumento, conformado fundamentalmente por escalas tipo Likert, identifica el grado de cumplimiento de las responsabilidades profesionales del docente y técnico docente que son inherentes a su profesión, su participación en el funcionamiento de la escuela, en órganos colegiados y su vinculación con los padres de familia y con la comunidad escolar; considerando la importancia de la Normalidad Mínima de Operación Escolar. El informe será emitido por el director de la escuela.

Expediente de evidencias de enseñanza

Este instrumento evalúa el análisis que realiza el docente y técnico docente sobre una selección de los productos de enseñanza de sus alumnos; dicho análisis contempla la descripción de las características del desarrollo y aprendizaje de los estudiantes; la explicación de las situaciones de aprendizaje que plantea, a partir del enfoque didáctico; la elección de los contenidos de aprendizaje para el logro de los propósitos educativos y la utilización de los resultados de la evaluación. También se valora la reflexión que realiza sobre su práctica y su vinculación con el aprendizaje de sus alumnos.

Evaluación de conocimientos actualizados y de las competencias didácticas que favorecen el aprendizaje y el logro de las competencias de los estudiantes

En el caso del docente se evalúan los conocimientos y las competencias didácticas por medio de dos exámenes; en el caso del técnico docente únicamente se evalúan las competencias didácticas, por medio de un examen:

Examen de conocimientos disciplinares¹

Este instrumento permite evaluar el dominio del conocimiento específico en la disciplina que corresponda, para favorecer el aprendizaje de los estudiantes.

El técnico docente no presentará este examen.

Examen de competencias didácticas

Este instrumento permite evaluar las competencias didácticas del docente y técnico docente para afrontar y resolver con éxito situaciones diversas de su práctica docente en Educación Media Superior.

Planeación didáctica argumentada

Este instrumento evalúa la capacidad del docente y técnico docente para analizar, justificar, sustentar y dar sentido a las estrategias de intervención didáctica elegidas para elaborar y desarrollar su planeación didáctica; así como la capacidad para analizar y reflexionar sobre lo que espera que aprendan sus estudiantes y sobre el uso de los resultados de las evaluaciones con fines de mejora.

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.

¹ Los docentes que imparten lengua adicional al español (inglés), deberán acreditar el dominio de esta lengua con la Certificación Nacional de Nivel de Idioma (CENNI, Nivel 12)

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:

- 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 deberán elaborar las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional. Para cada especificación deberá existir, al menos, un reactivo con el cual será evaluada.
- Los instrumentos de evaluación deberán tener, al menos, 80 reactivos efectivos para calificación y deberá documentarse el procedimiento que se siguió para determinar la estructura del instrumento y la cantidad de reactivos que lo conforman, a fin de justificar la relevancia (ponderación) de los contenidos específicos evaluados en el mismo.
- Para el diseño de los casos asociados a un único reactivo debe cuidarse la extensión, a fin de que se incorpore únicamente la información que sea indispensable para resolver el reactivo.
- Para el diseño de los casos con formato de multi-reactivo, deberá verificarse que: a) todos los reactivos necesiten del planteamiento general para ser contestados; b) los reactivos evalúen conocimientos o habilidades complejas, no de reconocimiento; c) los reactivos sean independientes entre sí, esto es, que para poder responderse no requieran de la información incorporada en alguno de ellos, o bien, de la respuesta dada a algún otro.

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 deberán elaborar las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional.
- Con base en las definiciones operacionales se diseñarán los niveles o categorías de ejecución que se incluirán en las rúbricas o guías de calificación.
- En las rúbricas o guías de calificación los distintos niveles o categorías de ejecución que se consignen, deberán ser claramente distinguibles entre sí y con un diseño ordinal ascendente (de menor a mayor valor).

Criterios y parámetros estadísticos

Los instrumentos empleados para la evaluación del desempeño deberán atender los siguientes criterios y parámetros estadísticos (Cook y Beckman 2006; Downing, 2004; Stemler y Tsai, 2008):

I. En el caso de los instrumentos de evaluación con reactivos de opción múltiple:

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

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 entre cada aspecto evaluado con la puntuación global deberá ser igual o mayor que 0.20.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.70.
- El porcentaje de acuerdos inter-jueces deberá ser 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. Estas mediciones deberán aportarse antes de emitir la calificación definitiva del sustentante, a fin de salvaguardar la confiabilidad de la decisión.

III. En el caso del Informe de cumplimiento de responsabilidades profesionales, para cada una de las escalas que lo constituyen:

- La correlación entre cada reactivo con la puntuación global de la escala deberá ser igual o mayor que 0.30.
- La confiabilidad del constructo medido a través de la escala deberá ser igual o mayor que 0.80.

Adicionalmente, para este instrumento, conformado fundamentalmente por escalas tipo Likert, se debe generar evidencia empírica de que los constructos se integran conforme a lo esperado, esto es: a) los reactivos se agrupan al o a los constructos previstos en el diseño del instrumento, es decir, la dimensionalidad de los reactivos corresponde a la estructura teórica de los constructos; b) hay una correlación positiva entre los distintos constructos que integran el instrumento; c) es factible implementar un modelo de medición a cada uno de los constructos, es decir, se cumplen los supuestos del modelo; y, d) el comportamiento diferencial que presentan los reactivos respecto a las poblaciones que se contrastan no perjudica a dichas poblaciones. (Muraki,1999; Wu y Adams, 2007; Bentler, 2006; Masters,1982).

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

En este apartado se utiliza la palabra “docentes” para hacer referencia tanto a docentes como a técnicos docentes.

Un paso crucial en el desarrollo y uso de los instrumentos de evaluación de naturaleza criterial, como es el caso de los que se utilizarán para 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 docente. En este sentido el estándar de desempeño delimita el marco interpretativo de las puntuaciones obtenidas en un instrumento por los sustentantes. El procedimiento para el establecimiento de puntos de corte y estándares de desempeño incluye tres fases, las cuales se describen a continuación:

Primera fase

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

Para todos los instrumentos se utilizarán 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 Expediente de evidencias de enseñanza

Niveles de desempeño	Descriptor
Nivel I (N I)	El docente presenta dificultades en el análisis y la explicación de su práctica y los efectos que ésta tiene en el aprendizaje de los estudiantes; y aunque ofrece argumentos sobre los métodos y materiales para impartir sus clases, se observan carencias en la comprensión de los elementos generales necesarios para llevar a cabo la evaluación de sus estudiantes de manera congruente con el marco normativo vigente.

Nivel II (N II)	El docente presenta elementos básicos en el análisis y la explicación sobre su práctica y los efectos que ésta tiene en el aprendizaje de los estudiantes, aunque ofrece argumentos sobre los métodos, materiales y diseño de estrategias de aprendizaje para impartir sus clases tomando en cuenta el entorno sociocultural y las necesidades de aprendizaje de sus alumnos. Asimismo, denota una comprensión adecuada de elementos generales requeridos para la evaluación de sus estudiantes, en concordancia con el marco normativo vigente.
Nivel III (N III)	El docente presenta el análisis completo y la explicación detallada de su práctica y los efectos que ésta tiene en el aprendizaje de los estudiantes. Asimismo, presenta argumentos adecuados sobre los métodos, materiales y diseño de estrategias de aprendizaje para impartir sus clases tomando en consideración el contexto, entorno sociocultural y las necesidades de aprendizaje de sus alumnos. En algunos casos utiliza los elementos de la evaluación de sus estudiantes para retroalimentar el proceso de enseñanza-aprendizaje, lo cual le permite aportar elementos generales acerca de su evaluación, en función del marco normativo vigente. Por otra parte, realiza acciones pedagógicas para contextualizar la importancia social, económica y política que tienen los temas de la asignatura que imparte.
Nivel IV (N IV)	El docente presenta el análisis y la explicación fundamentada sobre su práctica y los efectos que ésta tiene en el aprendizaje de los estudiantes, al mismo tiempo que presenta argumentos sólidos sobre los métodos, materiales y diseño de estrategias de aprendizaje sobre los cuales fundamenta sus acciones para impartir sus clases. En particular, tomando en cuenta el contexto, entorno sociocultural y la diversidad de necesidades de aprendizaje de sus alumnos y los retoma para problematizar la importancia social, económica y política que tienen los temas abordados en su asignatura. Por otra parte, en el marco normativo vigente, emplea con recurrencia los elementos de la evaluación de sus estudiantes para retroalimentar el proceso de enseñanza-aprendizaje y le permite también ofrecerles mayores elementos para establecer estrategias de aprendizaje que impulsen su creatividad, originalidad y enriquezcan su práctica cotidiana.

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 muestra carencia de conocimientos de la asignatura que imparte, así como de la organización del contenido teórico-metodológico y muestra dificultades para afrontar y resolver situaciones de su práctica docente para cumplir con el programa de estudios.
Nivel II (N II)	El docente muestra conocimientos básicos de la asignatura que imparte, aunque comprende la organización del contenido teórico-metodológico del programa de estudios, que le permiten afrontar y resolver diversas situaciones de su práctica docente. Por otra parte, demuestra pocas habilidades didácticas tales que le permitan favorecer en sus estudiantes el aprendizaje de los conocimientos disciplinares y trasladarlos a distintos escenarios o contextos.
Nivel III (N III)	El docente muestra conocimientos sólidos de la asignatura que imparte, así como de la organización del contenido teórico-metodológico del plan de estudios y sabe utilizarlos para afrontar y resolver diversas situaciones de su práctica docente, que faciliten el cumplimiento de objetivos de aprendizaje, así como el desarrollo de sus capacidades de traducir los conocimientos aprendidos de un contexto a otro.
Nivel IV (N IV)	El docente muestra conocimientos sólidos y organizados de la asignatura que imparte, así como una amplia comprensión del contenido teórico-metodológico del programa de estudios. Por otra parte, cuenta con una amplia capacidad para afrontar y resolver diversas situaciones de su práctica docente, cumpliendo con las metas establecidas de aprendizaje de sus estudiantes y enriqueciendo sus conocimientos, habilidades y destrezas dentro de distintos contextos y escenarios alternos de aprendizaje, fomentando en ellos la creatividad y originalidad.

Tabla 1c. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de competencias didácticas

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente muestra deficiencias para adaptar los conocimientos sobre la disciplina que imparte en su práctica docente, así como dificultades para planificar los procesos de enseñanza, aprendizaje y evaluación.
Nivel II (N II)	El docente adapta los conocimientos sobre la disciplina que imparte en su práctica docente de acuerdo con las características de los estudiantes; muestra habilidades básicas para planificar los procesos de enseñanza-aprendizaje y tiene noción sobre la construcción de ambientes de aprendizaje autónomo y colaborativo.
Nivel III (N III)	El docente adapta los conocimientos sobre la disciplina que imparte en su práctica docente de acuerdo con las características de los estudiantes, planifica adecuadamente los procesos de enseñanza-aprendizaje, ubicándolos en contextos curriculares y sociales; además de que construye ambientes de aprendizaje autónomo y colaborativo con base en el Marco Normativo y Ético.
Nivel IV (N IV)	El docente muestra amplia capacidad para adaptar los conocimientos sobre la disciplina que imparte en su práctica docente, de acuerdo con las características de los estudiantes y en pleno reconocimiento a su diversidad cultural. Asimismo, planifica eficientemente los procesos de enseñanza, aprendizaje y evaluación, ubicándolos en contextos curriculares y sociales, con una amplia capacidad de adaptación para construir ambientes de aprendizaje autónomo y colaborativo, basado en el Marco Normativo y Ético.

Tabla 1d. Descriptores genéricos de los niveles de desempeño para el instrumento Planeación didáctica argumentada²

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente demuestra dificultades para la planeación al analizar, justificar y orientar sus estrategias de intervención didáctica; organiza su práctica educativa a partir de elementos mínimos del contexto y presenta reflexión y argumentación acerca de elementos de su práctica docente, así como lo que espera que aprendan sus estudiantes y la forma en que evaluará lo aprendido.
Nivel II (N II)	El docente demuestra habilidades básicas para la planeación al analizar, justificar y orientar sus estrategias de intervención didáctica, organiza y aplica su práctica educativa a partir de los elementos del contexto y considera las características de sus estudiantes; asimismo, presenta reflexión y argumentación elemental acerca de su práctica docente y de herramientas de evaluación de los aprendizajes con base en lo que espera que aprendan sus estudiantes.
Nivel III (N III)	El docente demuestra habilidades y amplia capacidad para la planeación al analizar, justificar y orientar sus estrategias de intervención didáctica. Organiza y aplica su práctica educativa a partir de los elementos del contexto y las características de sus estudiantes; presenta una reflexión y argumentación completa acerca de su práctica docente y de herramientas de evaluación de los aprendizajes, así como lo que espera que aprendan sus estudiantes.
Nivel IV (N IV)	El docente demuestra gran capacidad para la planeación al analizar, justificar y orientar de forma innovadora sus estrategias de intervención didáctica para favorecer los aprendizajes de los estudiantes. Organiza y aplica su práctica educativa a partir de los elementos adecuados del contexto y las características de sus estudiantes; presenta una reflexión y argumentación minuciosa acerca de su práctica docente y de herramientas de evaluación de los aprendizajes, así como lo que espera que aprendan sus estudiantes.

² Los docentes que hayan participado en el proceso de Certificación de Competencias Docentes para la Educación Media Superior (CERTIDEMS), o bien, en el de Evaluación de Competencias Docentes para la Educación Media Superior (ECODEMS) y hayan obtenido dictamen favorable, podrán presentar la constancia correspondiente en lugar de sustentar la Planeación didáctica argumentada, pero deberán considerar que se les asignará un nivel de desempeño y un puntaje con base en las siguientes equivalencias:

- En el caso de haber presentado el certificado ECODEMS, se le asignará la puntuación menor del nivel desempeño N IV de Planeación didáctica argumentada.
- En el caso de haber presentado el certificado CERTIDEMS, la correspondencia es uno a uno, es decir, si se obtuvo N I en CERTIDEMS se le asignará N I en Planeación didáctica argumentada y así sucesivamente; el puntaje que recibirá será el puntaje medio del nivel de desempeño alcanzado.

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 y los ejemplos de tareas evaluativas o de reactivos incluidos en las mismas. 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 (N I, N II, N III o N IV). Para ello, se deberá determinar, para cada tarea evaluativa o reactivo considerado en el instrumento, cuál es la probabilidad de que dichos sustentantes hipotéticos lo respondan correctamente y, con base en la suma de estas probabilidades, establecer la calificación mínima requerida o punto de corte, para cada nivel de desempeño (Angoff, 1971).

Una vez establecidos los puntos de corte que dividen el rango de calificaciones para diferenciar los niveles de desempeño en cada instrumento, considerando el conjunto de reactivos que, en cada caso el sustentante hipotético es capaz de responder, se deberán describir los conocimientos y las habilidades específicos que están implicados en cada nivel de desempeño, en términos de lo que éste 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, una vez que se cuente con los resultados alcanzados por los sustentantes, a fin de determinar si es necesario realizar algún ajuste en la decisión tomada con anterioridad y, de ser el caso, llevar a cabo el ajuste correspondiente.

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

La tercera 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 los definidos 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 instrumento y resultado global

Todos los sustentantes que participen recibirán los resultados de cada uno de los instrumentos de evaluación³; así como el resultado global de todo el proceso de evaluación. A continuación, se presentan dos subapartados: en el primero se describen los procedimientos para calificar los resultados de los sustentantes en cada instrumento y la escala utilizada para reportarlos; en el segundo se detallan los procedimientos para la obtención del resultado global y la conformación de los grupos de desempeño.

4.1 Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen el proceso de evaluación y la escala en que se reportarán

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 contra el estándar previamente establecido, con independencia de los resultados obtenidos por el conjunto de sustentantes que presentaron el examen.

³ En el caso en que el sustentante **no presente alguno** de los instrumentos de evaluación que son considerados para la calificación global, 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 considerados para efectos de calificación, 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.

Proceso para la equiparación de instrumentos de evaluación

Cuando el proceso de evaluación implica la aplicación un instrumento en diversas ocasiones en un determinado periodo, en especial si sus resultados tienen un alto impacto, es indispensable el desarrollo y uso de formas o versiones 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

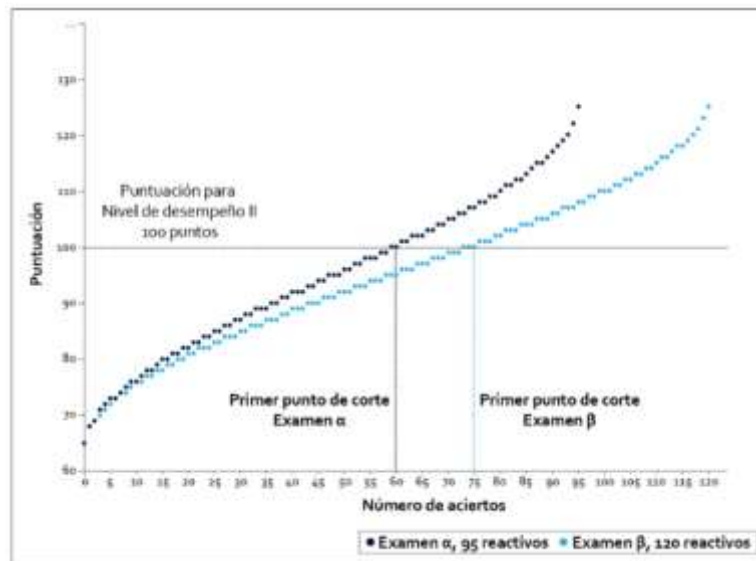
En cada plan de evaluación es indispensable definir la escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. En 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 este proceso de evaluación, permitirá construir una métrica común para todos los instrumentos que se administrarán. Consta de dos transformaciones, la primera denominada doble arcoseno, que permite estabilizar la magnitud de la precisión de las puntuaciones a lo largo de la escala; la segunda transformación es lineal y ubica el punto de corte del nivel de desempeño II en un mismo valor para todos los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos⁴).

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

En la siguiente gráfica puede observarse el número de aciertos obtenidos en dos instrumentos de longitudes diferentes y con puntos de corte distintos que, a partir del escalamiento, es posible graficar en una misma escala, trasladando el primer punto de corte a 100 puntos, aun cuando en cada 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 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.



4.2 Calificación del resultado global, escala en que se reportará y procedimiento para la conformación de los grupos de desempeño

El resultado global de la evaluación se realizará considerando los siguientes instrumentos:

Para los docentes

- o Expediente de evidencias de enseñanza
- o Examen de conocimientos disciplinares
- o Examen de competencias didácticas
- o Planeación didáctica argumentada

Para los técnicos docentes

- o Expediente de evidencias de enseñanza
- o Examen de competencias didácticas
- o Planeación didáctica argumentada

El **Informe de cumplimiento de responsabilidades profesionales** tiene por función contar con información sobre el cumplimiento del docente o técnico docente en sus funciones y es proporcionado por el directivo escolar que corresponda. Para fines de la calificación global, este instrumento no será considerado, por lo que la información que aporte el directivo escolar será utilizada sólo para efectos de diagnóstico y se dará retroalimentación al docente o técnico docente en el informe individual de resultados de la evaluación, sin afectar su calificación. En el caso de que el docente o técnico docente no tenga el informe, debido a que es responsabilidad de un tercero, se indicará que no se cuenta con la información.

La retroalimentación que recibirá el docente o técnico docente atenderá a los siguientes rubros, según el resultado del análisis del instrumento, a partir de la información disponible: participación en el trabajo escolar; trabajo colegiado; colaboración con la comunidad escolar y normalidad mínima en el desarrollo del ciclo escolar y del plantel.

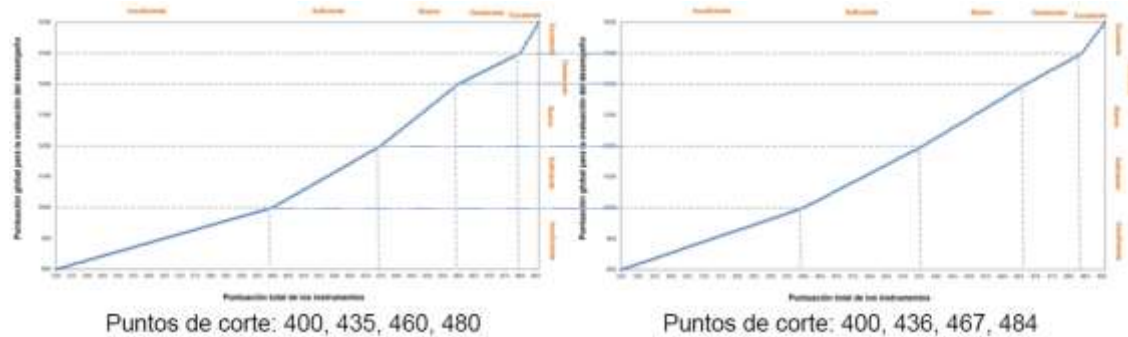
La calificación global

Una vez que se cuenta con las puntuaciones obtenidas en cada uno de los instrumentos de evaluación considerados para la calificación, se procede a calcular la puntuación total de cada sustentante como la suma de dichas puntuaciones. Esto tiene la ventaja de que los aspectos evaluados en que los docentes o técnicos docentes muestran mayores fortalezas compensan aquellos en los que son más débiles, es decir, la calificación está concebida como compensatoria.

Al sumar las puntuaciones obtenidas en cada instrumento de evaluación, los resultados presentan variabilidad, debido a que la escala es aplicable a múltiples instrumentos con características diversas dependiendo del campo y la asignatura evaluada. Por ello, es necesario transformar la suma de las distintas puntuaciones obtenidas a una escala global única de 800 a 1 600 puntos⁵. Esta operación permite homologar dichas puntuaciones, determinar puntos de corte para establecer los diferentes grupos de desempeño y comunicar de forma sencilla sus calificaciones a los sustentantes.

⁵ Para mayores detalles sobre el proceso que se lleva a cabo para la transformación de las puntuaciones a la escala global de 800 a 1 600 puntos, consultar el anexo.

A manera de ejemplo, en las siguientes gráficas se muestra la transformación de las puntuaciones en dos trayectos evaluativos diferentes en los que se consideran cuatro instrumentos para la calificación global; en el eje horizontal de ambas gráficas se muestra que la suma de las puntuaciones de los instrumentos, es decir, la puntuación total, será variable, esto se debe a que depende de la cantidad de instrumentos que se considere la evaluación y la puntuación alcanzada en cada uno de ellos. Sin embargo, la transformación que se realiza permitirá comunicar que para el resultado “Suficiente” se requieren, al menos, 1 000 puntos; para el grupo “Bueno”, al menos, 1 200 puntos, para el grupo “Destacado”, al menos, 1 400 puntos y para el grupo “Excelente”, al menos, 1 500 puntos (representados en el eje vertical).



Asimismo, en los ejemplos se observa que en ambas gráficas el punto de corte para alcanzar el resultado “Suficiente” es 400 puntos y la puntuación global está referida a 1 000 puntos. Para alcanzar el resultado global de “Bueno” se requieren 435 puntos en la gráfica de la izquierda y 436 puntos en la gráfica de la derecha, pero ambos están referidos a 1 200 puntos en la escala global. El resultado “Destacado” está referido a 1 400 puntos en la escala global en ambas gráficas, aun cuando en la gráfica de la izquierda se requieren 460 puntos y en la de la derecha 467 puntos. Esto mismo ocurre para el resultado “Excelente”, el cual está referido a 1 500 puntos en la escala global en ambas gráficas, aun cuando en la gráfica de la izquierda se requieren 480 puntos y en la de la derecha 484 puntos.

El resultado “Suficiente”

Criterios para que el docente o técnico docente obtenga un resultado Suficiente en el proceso de evaluación	
1)	Sustentar todos y cada uno de los instrumentos de la evaluación que son considerados para la calificación
2)	Obtener, al menos, el nivel de desempeño II (N II) en por lo menos tres de los cuatro instrumentos que integran el proceso de evaluación. En el caso de los técnicos docentes, en por lo menos dos de los tres instrumentos que integran el proceso de evaluación
3)	Obtener al menos 1 000 puntos en la escala de calificación global

Todos los instrumentos tendrán la misma jerarquía o peso en la puntuación total de la calificación, por lo que el resultado global de la evaluación del desempeño estará dado considerando la puntuación que aporta cada uno de los instrumentos. De esta forma, el efecto compensatorio de la puntuación global no se ve afectado ni distorsionado por una ponderación diferenciada de los instrumentos de evaluación.

La conformación de los grupos de desempeño

Los grupos de desempeño estarán conformados únicamente por los sustentantes que obtengan, al menos, un resultado “Suficiente” en la evaluación, esto es, hayan sustentado todos y cada uno de los instrumentos de la evaluación que son considerados para la calificación, alcanzado el criterio del número de instrumentos mínimos en N II y, al menos 1 000 puntos en la escala global. Con base en lo anterior, el resultado de la evaluación del desempeño del docente o técnico docente puede ubicarse en alguna de las siguientes categorías:

Resultado de la evaluación: grupos de desempeño

Criterios para formar parte de un grupo de desempeño en Educación Media Superior	
Resultado de la evaluación	Puntuación global en escala 800 - 1 600
Suficiente	De 1 000 a 1 199 puntos en la puntuación global
Bueno	De 1 200 a 1 399 puntos en la puntuación global
Destacado	De 1 400 a 1 499 puntos en la puntuación global
Excelente	Al menos 1 500 puntos en la puntuación global

El resultado “No se presentó a la evaluación”

Para el caso en que el docente o técnico docente no sustente NINGUNO de los instrumentos considerados para efectos de calificación, 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.

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 alguno o algunos de los instrumentos que son considerados para la calificación, los cuales son, para los docentes: i) Expediente de evidencias de enseñanza; ii) Examen de conocimientos disciplinares; iii) Examen de competencias didácticas y iv) Planeación didáctica argumentada; para los técnicos docentes: i) Expediente de evidencias de enseñanza; ii) Examen de competencias didácticas y iii) Planeación didáctica argumentada.
- **No obtenga**, al menos, N II en tres de los cuatro instrumentos, o en dos de los tres para el caso de los técnicos docentes.
- Haya obtenido **menos** de 1 000 puntos en el resultado global de la evaluación.

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

Se enfatiza que en cualquiera de estos escenarios los docentes y técnicos docentes recibirán los resultados alcanzados en los instrumentos de evaluación que hayan sustentado, a fin de proporcionarles retroalimentación para que conozcan sus fortalezas y áreas de oportunidad.

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 oportunidad
del personal docente en Educación Media Superior**

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, considerando los siguientes términos:

- Serán sujetos a una segunda oportunidad de evaluación del desempeño exclusivamente los Docentes que obtuvieron resultado insuficiente en su primera evaluación del desempeño. Estos docentes presentarán **únicamente** los instrumentos considerados para la calificación en los que haya obtenido nivel de desempeño I en su primera evaluación, así como aquellos a los que no se presentó (si fuera el caso).
- Para garantizar la equidad en los resultados de los docentes sujetos a una segunda oportunidad, se aplicará el mismo esquema de la primera evaluación del desempeño realizada en 2015, es decir, **sólo se considerarán los instrumentos que tuvieron efecto en la calificación** de los docentes en su primera oportunidad de evaluación.
- La calificación global se estimará considerando el resultado obtenido en los instrumentos de evaluación en los que obtuvo, al menos, nivel de desempeño II en su primera oportunidad, y se integrarán con la alcanzada en los instrumentos sustentados en esta segunda oportunidad, siguiendo los mismos principios del modelo de calificación desarrollado en los presentes criterios técnicos (véase la primera sección).
- Para los instrumentos que se sustenten en esta segunda oportunidad, sólo se deben considerar como referencia los puntos de corte empleados en el proceso de evaluación 2016-2017. Lo anterior debido a que los instrumentos de evaluación son sometidos a un proceso de mejora que implica una revisión de los puntos de corte, por lo que previsiblemente éstos tienen ligeros ajustes.

Los Docentes que obtengan un resultado insuficiente en la segunda oportunidad en la evaluación del desempeño, serán reincorporados a los programas de regularización que determine la Autoridad Educativa competente, para sujetarse a una tercera 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.

Anexo técnico

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

Protocolo de calificación por jueces para las rúbricas

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

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

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

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

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

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

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

- a. Se identifica la categoría asignada al sustentante en cada aspecto a evaluar.
- b. Se identifica el valor asignado a cada categoría de la rúbrica.
- c. La suma de los valores es el resultado de la calificación.

9. Las asignaciones de categorías del sustentante en cada aspecto a evaluar para emitir su calificación global 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 multiplicada por uno para el caso de reactivos -toda vez que cada reactivo vale un punto-; o bien, la suma de estas probabilidades multiplicada por el valor máximo posible de las categorías de la rúbrica). Las decisiones de los jueces se promedian obteniendo el punto de corte. La decisión del conjunto de jueces pasa por una primera ronda para valorar sus puntos de vista en plenaria y puede modificarse la decisión hasta llegar a un acuerdo en común.

Método de Beuk

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

Procedimiento

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

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

Los pasos a seguir son los siguientes:

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

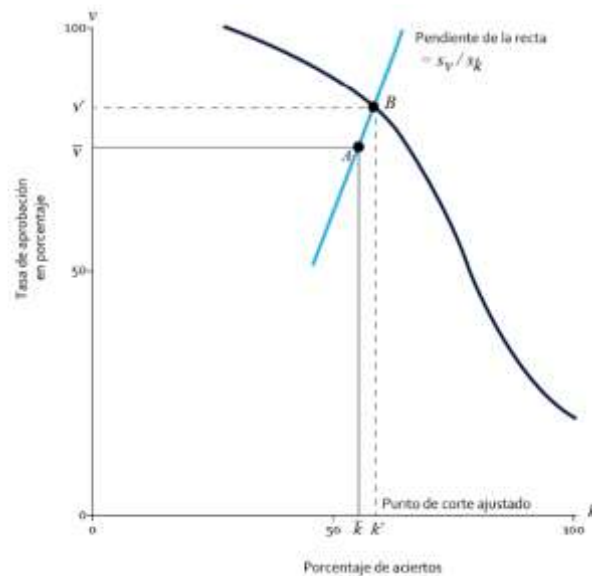
2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

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

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

El punto B, el cual tiene coordenadas (k', v') , representa los valores ya ajustados, por lo que k' corresponderá al punto de corte del estándar de desempeño.

El método asume que el grado en que los expertos están de acuerdo es proporcional a la importancia relativa que los expertos dan a las dos preguntas, de ahí que se utilice una línea recta con pendiente s_v/s_k .



Escalamiento de las puntuaciones

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

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

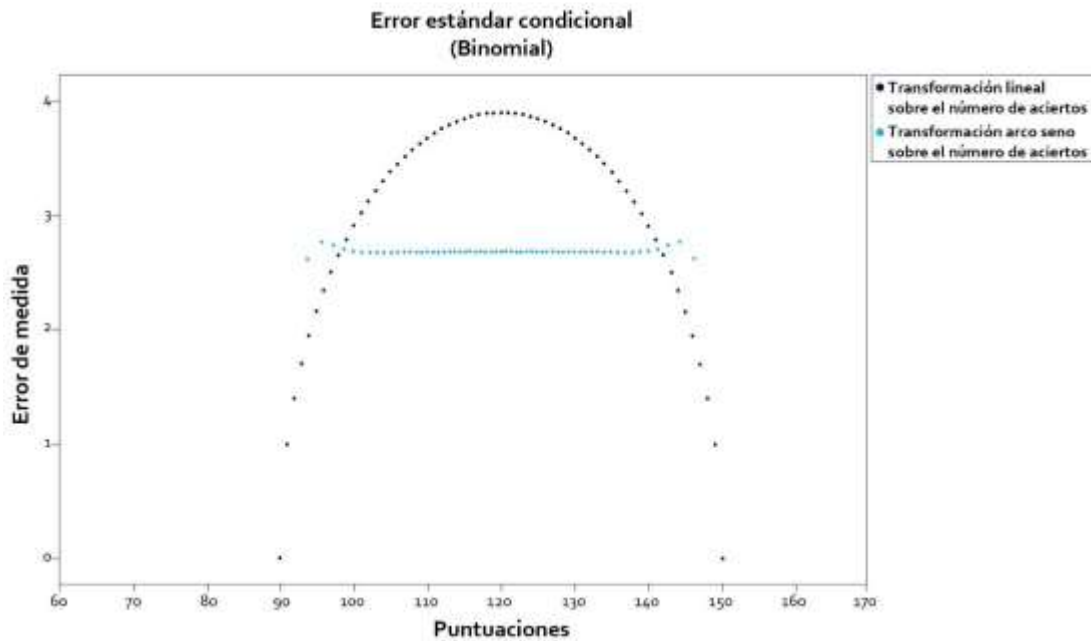
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

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 :

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

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:

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

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:

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

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:

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

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

Procedimiento para el error estándar condicional. Método delta

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

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

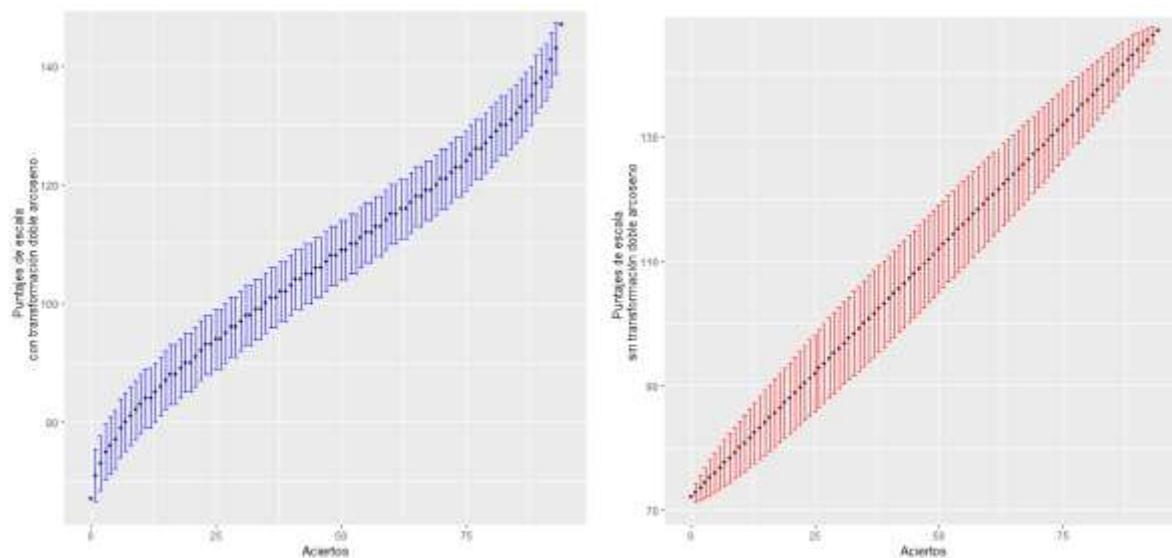
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

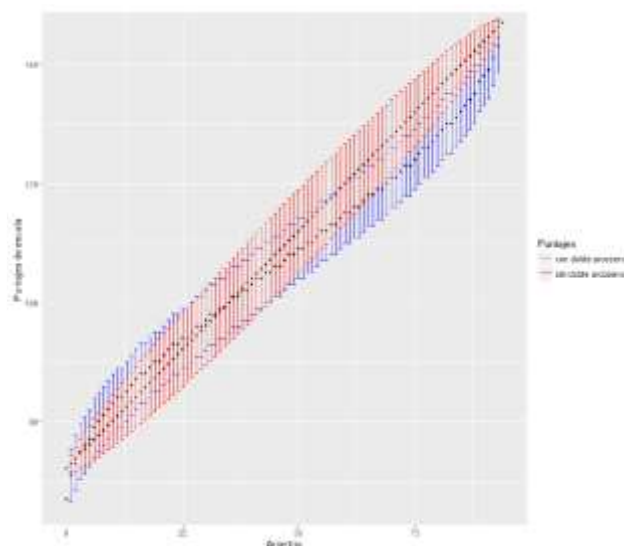
Donde $\sigma(x)$ es el error estándar de medida de las puntuaciones crudas y $\sigma(P_i)$ el error estándar condicional de medición, de la transformación P_i , que ya incorpora la confiabilidad.

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, $w_1 + w_2 = 1$ y $w_1, w_2 \geq 0$. Para este proceso se utilizará

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

y

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

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

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

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

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

$$I_{V_s}(x) = \frac{\sigma_s(V)}{\sigma_s(X)} [x - \mu_s(X)] + \mu_s(Y)$$

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 $P_{ie} = A + c(x_e) + B$, que ya incorpora la confiabilidad, está dado por:

$$\sigma(P_{ie}) \doteq \frac{A}{2} \left[\frac{b_1}{2(k+1) \left(\sqrt{\frac{x_e}{k+1}} \right) \left(\sqrt{1 - \frac{x_e}{k+1}} \right)} + \frac{b_1}{2(k+1) \left(\sqrt{\frac{x_e+1}{k+1}} \right) \left(\sqrt{1 - \frac{x_e+1}{k+1}} \right)} \right] \sigma(x_e)$$

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).

Algoritmo para el cálculo de la puntuación en escala global

En principio se calcula la puntuación total de los instrumentos, para el siguiente caso se consideran cuatro para efectos de calificación:

$$PT_i = \sum_{j=1}^{n_i} I_{ji}$$

PT_i = Puntuación total de los instrumentos que alcanza el sustentante i

I_{ji} = Puntuación que alcanza el sustentante i en el instrumento j

$j = 1, 2, 3, 4$ (Instrumento que forma parte de la evaluación [no se considera el informe de responsabilidades profesionales])

$n_i = 4$ (Número total de instrumentos que presenta el sustentante i)

Posteriormente, se establecen los cuatro puntos de corte globales considerando la escala de puntuaciones PT_i .

Finalmente, se calcula la puntuación en escala global, considerando los puntos de corte establecidos en el paso 2, los cuales serán asociados a 1 000, 1 200, 1 400 y 1 500 puntos respectivamente en la escala que va de 800 a 1 600 puntos.

$Si \min \{PT_i\} \leq PT_k < PC_1$	$G_k = 800 + \frac{(PT_k - \min\{PT_i\}) * 200}{PC_1 - \min\{PT_i\}}$
$Si PC_1 \leq PT_k < PC_2$	$G_k = 1\,000 + \frac{(PT_k - PC_1) * 200}{PC_2 - PC_1}$
$Si PC_2 \leq PT_k < PC_3$	$G_k = 1\,200 + \frac{(PT_k - PC_2) * 200}{PC_3 - PC_2}$
$Si PC_3 \leq PT_k < PC_4$	$G_k = 1\,400 + \frac{(PT_k - PC_3) * 100}{PC_4 - PC_3}$
$Si PC_4 \leq PT_k \leq \max \{PT_i\}$	$G_k = 1\,500 + \frac{(PT_k - PC_4) * 100}{\max \{PT_i\} - PC_4}$

G_k = Puntuación en escala global del sustentante k en la evaluación del desempeño

PT_k = Puntuación total que alcanza el sustentante k

$\max\{PT_i\}$ = Puntuación total máxima posible

$\min\{PT_i\}$ = Puntuación total mínima posible

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 treinta de septiembre de dos mil dieciséis.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil dieciséis, celebrada el treinta de septiembre de dos mil dieciséis. Acuerdo número SOJG/9-16/07, R. La Consejera Presidenta, **Sylvia Irene Schmelkes del Valle**.- Rúbrica.- Los Consejeros: **Eduardo Backhoff Escudero**, **Teresa Bracho González**, **Gilberto Ramón Guevara Niebla**, **Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 439159)

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, técnico docente y de quienes ejerzan funciones de dirección y supervisión en Educación Básica en el ciclo escolar 2016-2017, así como para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente en este mismo tipo educativo.

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, TÉCNICO DOCENTE Y DE QUIENES EJERZAN FUNCIONES DE DIRECCIÓN Y SUPERVISIÓN EN EDUCACIÓN BÁSICA EN EL CICLO ESCOLAR 2016-2017, ASÍ COMO PARA LLEVAR A CABO LA EVALUACIÓN DEL DESEMPEÑO EN SU SEGUNDA OPORTUNIDAD DEL PERSONAL DOCENTE EN ESTE MISMO TIPO EDUCATIVO.

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 III, incisos b), f) y h) y VI; 52, 53 y 54 de la Ley General del Servicio Profesional Docente; 22, 28, fracción X, 38, fracciones IX y XXII de la Ley del Instituto Nacional para la Evaluación de la Educación; de los Lineamientos para llevar a cabo la evaluación del desempeño del personal docente, técnico docente y de quienes ejerzan funciones de dirección y supervisión en educación básica y media superior en el ciclo escolar 2016-2017 (LINEE-09-2016), así como de los Lineamientos para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente y técnico docente en educación básica y media superior (LINEE-10-2016), 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, técnico docente y de quienes ejerzan funciones de dirección y supervisión en Educación Básica en el ciclo escolar 2016-2017, así como para llevar a cabo la evaluación del desempeño en su segunda oportunidad del personal docente en este mismo tipo educativo.

Los presentes Criterios técnicos y de procedimiento tienen como finalidad establecer los referentes necesarios para garantizar la validez, confiabilidad y equidad de los resultados de los procesos de evaluación. Su contenido se organiza de la siguiente manera:

Primera sección: Sobre la evaluación del desempeño para el ciclo escolar 2016-2017

Incorpora cuatro apartados: 1) Características generales de los instrumentos para evaluar el desempeño del personal docente, técnico docente y de quienes ejerzan 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; 4) Resultado de la evaluación del desempeño: resultado por instrumento y resultado global.

Segunda sección: Sobre la evaluación del desempeño de quienes será su segunda oportunidad, considerando sus resultados en el proceso de evaluación del desempeño para el ciclo escolar 2015-2016

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 del instrumento tienen consecuencias importantes para las personas o las instituciones; por ejemplo, en los procesos de admisión o certificación.
- II. **Calificación:** Proceso de asignación de una puntuación o nivel de desempeño logrado a partir de los resultados de una medición.
- III. **Confiabilidad:** Cualidad de las mediciones obtenidas con un instrumento, que se caracterizan por ser consistentes y estables cuando éste se aplica en distintas ocasiones.
- IV. **Constructo:** Elaboración teórica formulada para explicar un proceso social, psicológico o educativo y cuya adecuada descripción permite que sea susceptible de ser observable o medible.

- V. **Correlación punto biserial:** Medida de consistencia que se utiliza en el análisis de reactivos, indica si hay una correlación entre el resultado de un reactivo con el resultado global del examen.
- VI. **Criterio de evaluación:** Indicador de un valor aceptable sobre el cual se puede establecer o fundamentar un juicio de valor sobre el desempeño de una persona.
- VII. **Desempeño:** Resultado obtenido por el sustentante en un proceso de evaluación o en un instrumento de evaluación educativa.
- VIII. **Dificultad de un reactivo:** Indica la proporción de personas que responden correctamente el reactivo de un examen.
- IX. **Distractores:** Opciones de respuesta incorrectas del reactivo de opción múltiple, que probablemente serán elegidas por los sujetos con menor dominio en lo que se evalúa.
- X. **Dominio:** Conjunto de conocimientos, habilidades, destrezas, actitudes u otros atributos que tienen las siguientes propiedades: límites, extensión y definición. También se puede aplicar a contenidos, procedimientos u objetos.
- XI. **Educación básica:** Tipo de educación que comprende los niveles de preescolar, primaria y secundaria en todas sus modalidades, incluyendo la educación indígena, la especial y la que se imparte en los centros de educación básica para adultos.
- XII. **Educación media superior:** Tipo de educación que comprende el nivel de bachillerato, los demás niveles equivalentes a éste, así como la educación profesional que no requiere bachillerato o sus equivalentes.
- XIII. **Equiparación:** Proceso estadístico que se utiliza para ajustar las puntuaciones de las formas de un mismo instrumento, permite que las puntuaciones de una forma a otra sean utilizadas de manera intercambiable. La equiparación ajusta, por dificultad, las distintas formas que fueron construidas con contenidos y dificultad similar.
- XIV. **Error estándar de medida:** Desviación estándar de una distribución hipotética de errores de medida de una población.
- XV. **Escala:** Procedimiento para asignar números, puntuaciones o medidas 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 la calificación que obtienen los sustentantes en uno o varios instrumentos de evaluación, colocando las puntuaciones de los distintos instrumentos o formas a una escala común.
- XVII. **Especificaciones de tareas evaluativas o reactivos:** Descripción detallada de las características relevantes que se espera hagan los sujetos al sustentar el instrumento de evaluación y que es posible observar a través de las tareas evaluativas o los reactivos. Tienen el papel de guiar a los comités académicos en la elaboración y validación de las tareas evaluativas o de los reactivos y que éstos cuenten con los elementos necesarios para construirlos alineados al objeto de medida o constructo que se desea evaluar a través del instrumento.
- XVIII. **Estándar:** Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación y que son acordados por expertos en evaluación.
- XIX. **Evaluación:** Acción de emitir juicios de valor sobre un objeto, sujeto o evento que resultan de comparar los resultados de una medición u observación con un referente previamente establecido.
- 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:** Procedimiento de recolección de datos que suelen tener distintos formatos, atendiendo a la naturaleza de la evaluación, por ejemplo, instrumentos de selección de respuesta, instrumentos de respuesta construida, cuestionarios, observaciones, portafolios, entre otros.
- XXII. **Instrumento de evaluación referido a un criterio:** Instrumento que permite comparar el desempeño de las personas evaluadas, con un estándar preestablecido.

- XXIII. Jueceo:** Método en el cual se utiliza la opinión de expertos (denominados jueces) para determinar, entre otras cosas, la pertinencia de la validez de las tareas evaluativas o de los reactivos respecto a un dominio; el establecimiento de estándares de desempeño y puntos de corte; así como la calificación de reactivos de respuesta construida.
- 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 que sus propiedades puedan 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 un instrumento de evaluación, y que refiere a lo que la persona evaluada es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.
- 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 a 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 sujeto 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 sujetos que evalúa, coincide en la puntuación asignada al desempeño de los sujetos, 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 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.
- XXXIV. Puntuación:** Número de aciertos obtenidos en un instrumento de evaluación.
- XXXV. Reactivo:** Unidad básica de medida de un instrumento de evaluación que consiste en una pregunta o instrucción que requiere una respuesta del sujeto.
- 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 de un instrumento de evaluación de respuesta construida y que consiste en la ejecución de una actividad que es susceptible de 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.

Primera sección.

Evaluación del desempeño del personal docente, técnico docente y de quienes ejerzan funciones de dirección y supervisión en Educación Básica, 2016-2017

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

Personal docente y técnico docente*Informe de cumplimiento de responsabilidades profesionales*

Este instrumento, conformado fundamentalmente por escalas tipo Likert, identifica el grado de cumplimiento de las responsabilidades profesionales del docente o técnico docente que son inherentes a su profesión, su participación en el funcionamiento de la escuela, en órganos colegiados y su vinculación con los padres de familia y con la comunidad escolar, considerando la importancia de la Normalidad Mínima de Operación Escolar. El informe será emitido por el director de la escuela o, en su caso, por el supervisor de la Zona Escolar.

Expediente de evidencias de enseñanza

Este instrumento evalúa el análisis que realiza el docente o técnico docente sobre una selección de los productos de enseñanza de sus alumnos; dicho análisis contempla la descripción de las características del desarrollo y aprendizaje de los estudiantes; la explicación de las situaciones de aprendizaje que plantea, a partir del enfoque didáctico; la elección de los contenidos de aprendizaje para el logro de los propósitos educativos y la utilización de los resultados de la evaluación. También se valora la reflexión que realiza el docente o técnico docente sobre su práctica y su vinculación con el aprendizaje de sus alumnos.

Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos

A partir de la resolución de situaciones hipotéticas de la práctica educativa, este instrumento evalúa los conocimientos y las competencias didácticas que el docente o técnico docente pone en juego para propiciar el aprendizaje de los alumnos, la colaboración en la escuela y el vínculo con los padres de familia y la comunidad.

Planeación didáctica argumentada

Este instrumento evalúa la capacidad del docente o técnico docente para analizar, justificar, sustentar y dar sentido a las estrategias de intervención didáctica elegidas para elaborar y desarrollar su planeación didáctica; así como la capacidad para analizar y reflexionar sobre lo que espera que aprendan sus alumnos y sobre el uso de los resultados de las evaluaciones con fines de mejora.

Examen complementario

Este instrumento evalúa el nivel de dominio específico en un campo de conocimiento, a saber: En secundaria: Segunda lengua: inglés; Segunda lengua: francés; y Tecnologías. En Educación Preescolar y Primaria Indígena: Lengua indígena.

Personal con funciones de dirección*Expediente de evidencias de la función de dirección*

Este instrumento evalúa el análisis y la reflexión que el director hace de la información contenida en las evidencias de su gestión directiva, a partir de argumentar las decisiones y acciones que toma en el ejercicio de su función.

Examen de conocimientos y habilidades directivas

Este instrumento evalúa los conocimientos y las competencias que el director pone en juego para propiciar el adecuado funcionamiento de la escuela al coordinar, gestionar y ejercer un liderazgo en los procesos de mejora continua, a partir de la resolución de situaciones hipotéticas de la práctica escolar.

Ruta de mejora argumentada

Este instrumento evalúa la capacidad del director para analizar el contexto interno y externo de la escuela, jerarquizar prioridades y definir metas en la Ruta de mejora, así como la argumentación que hace para propiciar la mejora continua de la escuela y los aprendizajes de los alumnos.

Personal con funciones de supervisión*Expediente de evidencias de la función de supervisión*

Este instrumento evalúa el análisis y la reflexión que el personal con funciones de supervisión hace de la información contenida en las evidencias de su gestión, a partir de argumentar las decisiones y acciones que toma en el ejercicio de su función.

Examen de conocimientos y habilidades de las funciones de supervisión

Este instrumento evalúa los conocimientos y las competencias que el personal con funciones de supervisión pone en juego para propiciar el adecuado funcionamiento de las escuelas al coordinar, gestionar y ejercer un liderazgo en los procesos de mejora continua, a partir de la resolución de situaciones hipotéticas de gestión en la supervisión.

Plan de trabajo argumentado

Este instrumento evalúa la capacidad del supervisor, jefe de sector o jefe de enseñanza para analizar el contenido y la estructura del Plan de trabajo, así como el diagnóstico y la descripción del contexto interno y externo de las escuelas, así como la argumentación que hace sobre las prioridades que establece para propiciar la mejora continua de las escuelas y los aprendizajes de los alumnos

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:

- 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 deberán elaborar las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional. Para cada especificación deberá existir, al menos, un reactivo con el cual será evaluada.
- Los instrumentos de evaluación deberán tener, al menos, 80 reactivos efectivos para calificación y deberá documentarse el procedimiento que se siguió para determinar la estructura del instrumento y la cantidad de reactivos que lo conforman, a fin de justificar la relevancia (ponderación) de los contenidos específicos evaluados en el mismo.
- Para el diseño de los casos asociados a un único reactivo debe cuidarse la extensión, a fin de que se incorpore únicamente la información que sea indispensable para resolver el reactivo.
- Para el diseño de los casos con formato de multi-reactivo, deberá verificarse que: a) todos los reactivos necesiten del planteamiento general para ser contestados; b) los reactivos evalúen conocimientos o habilidades complejas, no de reconocimiento; c) los reactivos sean independientes entre sí, esto es, que para poder responderse no requieran de la información incorporada en alguno de ellos, o bien, de la respuesta dada a algún otro.

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 deberán elaborar las especificaciones de las tareas evaluativas. Cada especificación deberá tener su definición operacional.
- Con base en las definiciones operacionales se diseñarán los niveles o categorías de ejecución que se incluirán en las rúbricas o guías de calificación.
- En las rúbricas o guías de calificación los distintos niveles o categorías de ejecución que se consignen, deberán ser claramente distinguibles entre sí y con un diseño ordinal ascendente (de menor a mayor valor).

Criterios y parámetros estadísticos

Los instrumentos empleados para la evaluación del desempeño deberán atender los siguientes criterios y parámetros estadísticos (Cook y Beckman 2006; Downing, 2004; Stemler y Tsai, 2008):

I. En el caso de los instrumentos de evaluación con reactivos de opción múltiple:

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

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 entre cada aspecto evaluado con la puntuación global deberá ser igual o mayor que 0.20.
- La confiabilidad del instrumento deberá ser igual o mayor que 0.70.
- El porcentaje de acuerdos inter-jueces deberá ser 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. Estas mediciones deberán aportarse antes de emitir la calificación definitiva del sustentante, a fin de salvaguardar la confiabilidad de la decisión.

III. En el caso del Informe de cumplimiento de responsabilidades profesionales, para cada una de las escalas que lo constituyen:

- La correlación entre cada reactivo con la puntuación global de la escala deberá ser igual o mayor que 0.30.
- La confiabilidad del constructo medido a través de la escala deberá ser igual o mayor que 0.80.

Adicionalmente, para este instrumento, conformado fundamentalmente por escalas tipo Likert, se debe generar evidencia empírica de que los constructos se integran conforme a lo esperado, esto es: a) los reactivos se agrupan al o a los constructos previstos en el diseño del instrumento, es decir, la dimensionalidad de los reactivos corresponde a la estructura teórica de los constructos; b) hay una correlación positiva entre los distintos constructos que integran el instrumento; c) es factible implementar un modelo de medición a cada uno de los constructos, es decir, se cumplen los supuestos del modelo; y, d) el comportamiento diferencial que presentan los reactivos respecto a las poblaciones que se contrastan no perjudica a dichas poblaciones. (Muraki, 1999; Wu y Adams, 2007; Bentler, 2006; Masters, 1982).

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

Un paso crucial en el desarrollo y uso de los instrumentos de evaluación de naturaleza criterial, como es el caso de los que se utilizarán para 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 para el caso de los docentes y técnicos docentes (para estos últimos no aplica la tabla 1d).

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

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente ofrece evidencia que denota falta de claridad en su práctica de enseñanza, así como dificultades para ajustar su intervención docente en función de las características del contexto de sus alumnos. Por otra parte, aunque presenta argumentos respecto a la elección de los contenidos de aprendizaje por desarrollar en su intervención docente, éstos son frágiles o poco consistentes con los propósitos educativos.
Nivel II (N II)	El docente muestra evidencias de su práctica de enseñanza a través de las cuales denota que realiza las acciones necesarias para ajustar su intervención docente en función de las características del contexto de sus alumnos. Además, presenta argumentos esenciales de la elección de los contenidos de aprendizaje a desarrollar en su intervención docente, aunque con parcial correspondencia con los propósitos educativos. Asimismo, se observan evidencias indispensables de la utilización de los resultados de la evaluación de sus alumnos para mejorar su práctica docente.
Nivel III (N III)	El docente aporta evidencias de su práctica de enseñanza que denotan un adecuado trabajo con sus alumnos, a través de su reflexión. Por otra parte, realiza acciones relevantes para ajustar su intervención docente en función de las características del contexto de sus alumnos y argumenta la elección de los contenidos de aprendizaje a desarrollar en su intervención docente, además de que los alinea con el objetivo de alcanzar los propósitos educativos planteados. También se observan evidencias satisfactorias de la utilización de los resultados de la evaluación de sus alumnos para mejorar su práctica docente.
Nivel IV (N IV)	El docente demuestra evidencias sólidas de su práctica de enseñanza y denota, a través de su proceso de reflexión, un destacado trabajo con sus alumnos en el aula, al ser sensible a la diversidad y complejidad de las características de sus estudiantes. También realiza acciones y estrategias óptimas para ajustar su intervención docente en función de las características del contexto del entorno en que se desarrolla el acto educativo y argumenta de forma clara y detallada la elección de los contenidos de aprendizaje a desarrollar en su intervención docente, con plena correspondencia con lograr los propósitos educativos planteados. Por otra parte, se observan evidencias notables de la utilización de los resultados de la evaluación de sus alumnos, que le permiten retroalimentar sus fortalezas y áreas de oportunidad, tales que las acciones que realiza son consistentes con un proceso de mejora de su práctica docente, incorporando elementos creativos que le permiten enriquecerla.

Tabla 1b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente muestra carencia de conocimientos para atender situaciones de la práctica docente; presenta dificultades para organizar el trabajo educativo y su intervención didáctica; así como para identificar los conceptos de los principios filosóficos, los fundamentos legales y las finalidades de la educación mexicana en el ejercicio de su función docente. Asimismo, muestra escasos conocimientos sobre las estrategias elementales que pueden contribuir a la construcción de un ambiente favorable en el entorno escolar.
Nivel II (N II)	El docente demuestra conocimientos básicos para atender situaciones de la práctica docente, así como habilidades para establecer vínculos con la comunidad en la que se encuentra la escuela. Sin embargo, organiza el trabajo educativo y su intervención didáctica e identifica los principios filosóficos, los fundamentos legales y las finalidades de la educación mexicana en el ejercicio de su función docente. Asimismo, distingue las diferentes estrategias didácticas que le pueden posibilitar el que contribuya a la construcción de un ambiente favorable en el entorno escolar.
Nivel III (N III)	El docente demuestra conocimientos sólidos para resolver situaciones de la práctica docente, así como habilidades para establecer vínculos con la comunidad en la que se encuentra la escuela, identificando el contexto en el cual está inmersa. Por otra parte, organiza de una manera eficaz el trabajo educativo y su intervención didáctica, además de comprender adecuadamente los principios filosóficos, los fundamentos legales y las finalidades de la educación mexicana y comprende la relevancia de implementarlos en el ejercicio de su función docente. Asimismo, incorpora estrategias para el establecimiento de ambientes favorables para la sana convivencia y la inclusión educativa en su práctica docente.
Nivel IV (N IV)	El docente demuestra conocimientos y habilidades sólidas y organizadas que aplica para resolver situaciones de la práctica educativa, que le demandan creatividad y originalidad. Además, organiza y sistematiza de una manera óptima el trabajo educativo y su intervención didáctica en el aula, con el reconocimiento de la diversidad cultural de sus alumnos; además analiza los principios filosóficos, los fundamentos legales y las finalidades de la educación mexicana y, con base en ellos implementa y promueve distintas actividades pedagógicas en el ejercicio de su función docente. Asimismo, construye ambientes favorables para la sana convivencia y la inclusión educativa de sus alumnos en su práctica docente y participa en el funcionamiento eficaz de la escuela, contribuyendo activamente a fomentar vínculos estrechos con la comunidad. Además, realiza acciones en la gestión escolar que aportan elementos que contribuyen a una mejora en la calidad de los resultados educativos de sus alumnos, a través del análisis, la reflexión y el trabajo colaborativo.

Tabla 1c. Descriptores genéricos de los niveles de desempeño para el instrumento Planeación didáctica argumentada

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente presenta dificultades para organizar, analizar, justificar, sustentar y adecuar estrategias para argumentar su intervención didáctica, describe técnicas con las que evaluará a sus alumnos, pero no explicita el propósito educativo. Demuestra reflexión limitada sobre su práctica docente, las características de su contexto y los efectos que éstas tienen en el aprendizaje de sus alumnos, identifica parcialmente el currículo vigente, denotando un ejercicio de análisis elemental acerca de lo que espera que aprendan sus alumnos. Asimismo, muestra conocimientos limitados de las estrategias de evaluación para una intervención didáctica.

Nivel II (N II)	El docente demuestra habilidades básicas para organizar, analizar, justificar, sustentar y adecuar estrategias para argumentar su intervención didáctica y aunque describe técnicas con las que evaluará a sus alumnos, la explicación del propósito educativo es limitada. Por otra parte, la reflexión sobre su práctica docente es escasa y la comprensión de las características de su contexto y los efectos que éstas tienen en el aprendizaje de sus alumnos no cobran la relevancia esperada en su argumentación. Sin embargo, identifica el currículo vigente para organizar su intervención docente y realiza un ejercicio de análisis acerca de lo que espera que aprendan sus alumnos y muestra los conocimientos necesarios sobre las estrategias de evaluación para la realización de una intervención didáctica.
Nivel III (N III)	El docente demuestra un dominio sólido de las habilidades necesarias para organizar y adecuar estrategias para argumentar su intervención didáctica; describe técnicas y métodos con los que evaluará a sus alumnos explicando el propósito educativo. Demuestra reflexión sobre su práctica docente, las características de su contexto y los efectos que éstas tienen en el aprendizaje de sus alumnos. Por otra parte, identifica el currículo vigente para organizar su intervención docente, denotando un adecuado ejercicio de análisis acerca de lo que espera que aprendan sus alumnos, así como una apropiada comprensión de las características y procesos de aprendizaje de los alumnos. Asimismo, muestra conocimientos amplios de las estrategias de evaluación y propone las que son pertinentes para su intervención didáctica específica en el aula.
Nivel IV (N IV)	El docente manifiesta amplias habilidades para organizar, analizar, justificar, sustentar y adecuar estrategias que le permiten aportar sólidos argumentos que fundamentan y orientan su intervención didáctica; selecciona las técnicas y métodos más adecuados al tipo de evaluación que llevará con sus alumnos, explicando el propósito educativo y los alcances, limitaciones y áreas de oportunidad que se puedan derivar. Demuestra una amplia reflexión sobre su práctica docente, las características del contexto interno y externo de la escuela y los efectos que éstas tienen en el aprendizaje de sus alumnos; identifica el currículo vigente para organizar su intervención docente, denotando un ejercicio de análisis acerca de lo que espera que aprendan sus alumnos y, con base en el diagnóstico de sus características y procesos de aprendizaje, propone las estrategias de evaluación más adecuadas y pertinentes para una intervención didáctica eficaz en el aula, adaptada a las condiciones del contexto y sensibles a las distintas necesidades de aprendizaje de sus alumnos.

Tabla 1d. Descriptores genéricos de los niveles de desempeño para uno de los instrumentos complementarios: Examen complementario. Segunda lengua: inglés.

Nivel de desempeño	Descriptor
Nivel I (N I)	El docente posee dominio limitado del idioma inglés; carece de capacidad para comprender textos extensos, así como ideas abstractas. Asimismo, muestra dificultades para comprender discursos de temas complejos, pronunciados en un nivel de lengua estándar que pueden incluir argumentaciones; sin embargo, puede distinguir en textos el significado de palabras e ideas principales.
Nivel II (N II)	El docente posee un dominio básico del idioma inglés; tiene capacidad limitada para comprender textos extensos, reconocer ideas abstractas, así como para comprender discursos de temas complejos, pronunciados en un nivel de lengua estándar que pueden incluir argumentaciones.
Nivel III (N III)	El docente posee un dominio intermedio del idioma inglés; es capaz de comprender textos extensos, así como de reconocer en ellos sentidos implícitos y términos técnicos. Demuestra habilidades para comprender discursos de temas complejos, pronunciados en un nivel de lengua estándar que pueden incluir argumentaciones.
Nivel IV (N IV)	El docente posee un dominio avanzado del idioma inglés; es capaz de comprender una amplia variedad de textos extensos, así como reconocer en ellos sentidos implícitos, ideas abstractas y términos técnicos. Demuestra altas habilidades para comprender discursos de temas complejos, pronunciados en un nivel de lengua estándar que pueden incluir argumentaciones complejas.

Para el caso de quienes ejercen funciones de dirección, los descriptores genéricos para los diferentes grupos de instrumentos y cada nivel se indican en las Tablas 2a, 2b y 2c

Tabla 2a. Descriptores genéricos de los niveles de desempeño para el instrumento Expediente de evidencias de la función de dirección

Nivel de desempeño	Descriptor
Nivel I (N I)	El directivo, aunque presenta evidencias de sus acciones para organizar el trabajo de la escuela, los argumentos que aporta sobre las decisiones y acciones tomadas en conjunto con el Consejo Técnico Escolar y el Consejo Escolar de Participación Social son faltos de sistematización y estructura, por lo que no se observa una clara relación con el objetivo de mejora de la calidad de los aprendizajes.
Nivel II (N II)	El directivo presenta evidencias básicas de sus acciones para organizar el trabajo de la escuela, y aporta argumentos que justifican las decisiones y acciones tomadas en conjunto con el Consejo Técnico Escolar y el Consejo Escolar de Participación Social y que han sido encaminadas a lograr el objetivo de la mejora de la calidad de los aprendizajes. Asimismo, ofrece una adecuada reflexión sobre sus cualidades y aptitudes en el ejercicio de su función.
Nivel III (N III)	El directivo presenta sólidas evidencias de sus acciones para organizar el trabajo de la escuela, y aporta argumentos relevantes que justifican las decisiones y acciones tomadas en conjunto con el Consejo Técnico Escolar y el Consejo Escolar de Participación Social y que han sido encaminadas a lograr el objetivo de la mejora de la calidad de los aprendizajes, tomando en consideración el contexto de la escuela, la priorización de objetivos, en función de las necesidades más apremiantes. Asimismo, ofrece una adecuada reflexión sobre sus cualidades y aptitudes en el ejercicio de su función, así como la necesidad de continuar con su desarrollo profesional.
Nivel IV (N IV)	El directivo presenta sólidas evidencias de sus acciones para organizar el trabajo de la escuela, y aporta, con exhaustividad, argumentos relevantes que justifican las decisiones y acciones tomadas en conjunto con el Consejo Técnico Escolar y el Consejo Escolar de Participación Social y que han sido encaminadas a lograr el objetivo de la mejora de la calidad de los aprendizajes. Se observa también que consideró el contexto de la escuela, el entorno socio-cultural y otorgó especial atención a las características y a la diversidad de los estudiantes; a partir de ello y de las necesidades identificadas, se observa que se priorizaron estratégicamente los objetivos y las posibilidades de intervención y gestión, en colaboración con las familias, la comunidad e incluso de otras instituciones interesadas en la mejora educativa de la escuela. Por otra parte, ofrece una adecuada reflexión sobre sus cualidades y aptitudes en el ejercicio de su función, así como la necesidad de enriquecer su desarrollo profesional, con la disponibilidad de realizarlo con apoyo de los recursos tecnológicos disponibles.

Tabla 2b. Descriptores genéricos de los niveles de desempeño para el instrumento Examen de conocimientos y habilidades directivas

Nivel de desempeño	Descriptor
Nivel I (N I)	El directivo muestra carencia de conocimientos para la adecuada organización y funcionamiento escolar, así como limitados conocimientos de los componentes del currículo, del aprendizaje de los alumnos, de los elementos del trabajo en el aula y de las prácticas docentes. Asimismo, presenta carencias fundamentales respecto a la relevancia de incorporar en la gestión estrategias para una adecuada construcción de ambientes de inclusión y equidad en la escuela.
Nivel II (N II)	El directivo demuestra conocimientos básicos sobre las formas de organización y funcionamiento escolar, así como de los componentes del currículo, del aprendizaje de los alumnos, de los elementos del trabajo en el aula y de las prácticas docentes que le permitan favorecer, a través de su desempeño, el funcionamiento eficaz de la escuela. También, a través de su gestión, incorpora prácticas que promueven la construcción de ambientes de inclusión y equidad viables en situaciones escolares, así como la promoción de principios éticos, filosóficos y fundamentos legales en contextos educativos.

Nivel III (N III)	El directivo demuestra conocimientos y habilidades sólidas para realizar diversas formas de organización y funcionamiento escolar, considerando las situaciones de contexto y diversidad de la comunidad escolar y del entorno socio-cultural en el cual se desenvuelve el acto educativo. Asimismo, demuestra un dominio sólido de los distintos componentes del currículo, del aprendizaje de los alumnos, de los elementos del trabajo en el aula y de las prácticas docentes que le permitan favorecer, a través de su desempeño, el funcionamiento eficaz de la escuela. También, a través de su gestión, incorpora prácticas que promueven la construcción de ambientes de inclusión, equidad y respeto, viables en situaciones escolares, así como la promoción de principios éticos, filosóficos y fundamentos legales en contextos educativos y adaptarlos a las demandas específicas de las condiciones escolares y de la población estudiantil. Además, analiza su práctica profesional para mejorarla.
Nivel IV (N IV)	El directivo demuestra conocimientos y habilidades sólidas y sistematizadas para implementar diversas formas de organización y funcionamiento escolar, con amplia sensibilidad a las situaciones de contexto de la comunidad escolar y del entorno socio-cultural en el cual se desenvuelve el acto educativo. Además, demuestra un amplio dominio de los distintos componentes del currículo, del aprendizaje de los alumnos, de los elementos del trabajo en el aula y de las prácticas docentes, tales que le permitan favorecer, a través de su desempeño, el funcionamiento eficiente y eficaz de la escuela. Por otra parte, gestiona y promueve la construcción de ambientes de inclusión, equidad, respeto y reconocimiento a la diversidad en situaciones escolares, así como de la promoción de principios éticos, filosóficos y fundamentos legales en contextos educativos. También realiza gestiones para contribuir a la mejora educativa y a la calidad de los aprendizajes de los alumnos al establecer vínculos entre el trabajo de la escuela y el de la zona escolar. Por otra parte, reflexiona sistemáticamente sobre su práctica profesional para mejorarla, a través de mecanismos de autoaprendizaje, formación y capacitación profesional, con el uso de las tecnologías de la información.

Tabla 2c. Descriptores genéricos de los niveles de desempeño para el instrumento Ruta de mejora argumentada

Nivel de desempeño	Descriptor
Nivel I (N I)	El director muestra dificultades para plantear cuáles son los elementos que integran la Ruta de mejora, además de que presenta debilidades para analizar el contexto interno y externo de la escuela, hacer un diagnóstico de la misma, jerarquizar prioridades y definir metas para el plantel, así como para evaluar los resultados de su gestión.
Nivel II (N II)	El director demuestra conocimientos y habilidades básicas para argumentar los elementos indispensables que integran la Ruta de mejora, sin embargo, muestra capacidad para analizar el contexto interno y externo de la escuela, hacer un diagnóstico de la misma, jerarquizar prioridades y definir metas para el plantel, así como para evaluar los resultados de su gestión, tales que le permitan plantear acciones encaminadas hacia la mejora continua y el funcionamiento de la escuela, cumpliendo con la Normalidad Mínima de Operación Escolar.
Nivel III (N III)	El director demuestra conocimientos y habilidades sólidas para argumentar, justificar y dar sustento a los elementos que integran la Ruta de mejora, además de que muestra capacidades para analizar el contexto interno y externo de la escuela, hacer un diagnóstico de la misma, jerarquizar prioridades y definir metas para el plantel. Por otra parte, establece acciones de vinculación entre la escuela y la comunidad para fortalecer los aprendizajes de los alumnos y evalúa los resultados de su gestión para plantear acciones encaminadas hacia la mejora continua y el funcionamiento eficaz de la escuela, cumpliendo con la Normalidad Mínima de Operación Escolar.
Nivel IV (N IV)	El director demuestra conocimientos y habilidades sólidas y organizadas para argumentar, justificar y dar sustento a los elementos que integran la Ruta de mejora. Además, muestra amplia capacidad para analizar el contexto interno y externo de la escuela, hacer un diagnóstico de la situación y con base en ello, establecer prioridades y metas para el plantel, buscando la vinculación entre la escuela y la comunidad a fin de fortalecer los aprendizajes de los alumnos. Por otra parte, evalúa los resultados de su gestión para plantear acciones encaminadas hacia la mejora continua y el funcionamiento eficiente y eficaz de la escuela, haciendo uso responsable de los recursos con los que cuenta y cumpliendo con la Normalidad Mínima de Operación Escolar. Además, diseña estrategias viables para la atención a los alumnos en riesgo de reprobación, rezago o deserción escolar, incorporando en ellas a los padres de familia, a la planta docente y a la comunidad escolar en general.

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 y los ejemplos de tareas evaluativas o de reactivos incluidos en las mismas. 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 (N I, N II, N III o N IV). Para ello, se deberá determinar, para cada tarea evaluativa o reactivo considerado en el instrumento, cuál es la probabilidad de que dichos sustentantes hipotéticos lo respondan correctamente y, con base en la suma de estas probabilidades, establecer la calificación mínima requerida o punto de corte, para cada nivel de desempeño (Angoff, 1971).

Una vez establecidos los puntos de corte que dividen el rango de calificaciones para diferenciar los niveles de desempeño en cada instrumento, considerando el conjunto de reactivos que, en cada caso el sustentante hipotético es capaz de responder, se deberán describir los conocimientos y las habilidades específicos que están implicados en cada nivel de desempeño, en términos de lo que éste 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, una vez que se cuente con los resultados alcanzados por los sustentantes, a fin de determinar si es necesario realizar algún ajuste en la decisión tomada con anterioridad y, de ser el caso, llevar a cabo el ajuste correspondiente.

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

La tercera 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 los definidos 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 instrumento y resultado global

Todos los sustentantes que participen recibirán los resultados de cada uno de los instrumentos de evaluación¹; así como el resultado global de todo el proceso de evaluación. A continuación, se presentan dos subapartados: en el primero se describen los procedimientos para calificar los resultados de los sustentantes en cada instrumento y la escala utilizada para reportarlos; en el segundo se detallan los procedimientos para la obtención del resultado global y la conformación de los grupos de desempeño.

4.1 Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen el proceso de evaluación y la escala en que se reportarán

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 contra el estándar previamente establecido, con independencia de los resultados obtenidos por el conjunto de sustentantes que presentaron el examen.

¹ En el caso en que el sustentante **no presente alguno** de los instrumentos de evaluación que son considerados para la calificación global, 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 considerados para efectos de calificación, 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.

Proceso para la equiparación de instrumentos de evaluación

Cuando el proceso de evaluación implica la aplicación un instrumento en diversas ocasiones en un determinado periodo, en especial si sus resultados tienen un alto impacto, es indispensable el desarrollo y uso de formas o versiones 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

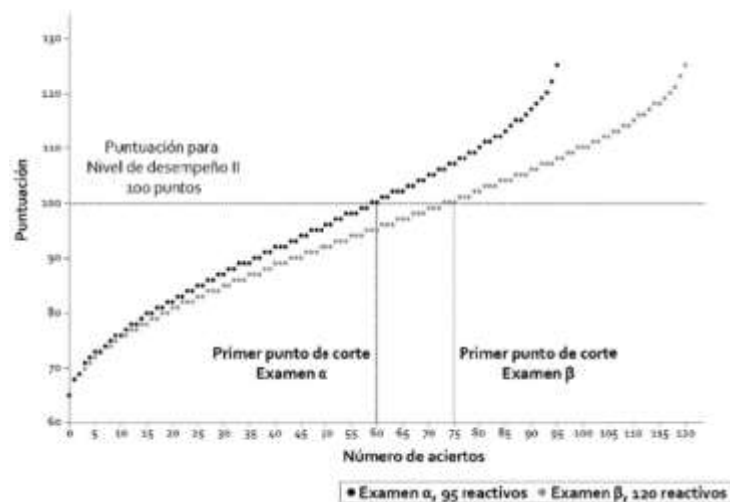
En cada plan de evaluación es indispensable definir la escala en la que se reportarán los resultados de los sustentantes. Existen muchos tipos de escalas de calificación; en las escalas referidas a norma, las calificaciones indican la posición relativa del sustentante en una determinada población. En 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 este proceso de evaluación, permitirá construir una métrica común para todos los instrumentos que se administrarán. Consta de dos transformaciones, la primera denominada doble arcoseno, que permite estabilizar la magnitud de la precisión de las puntuaciones a lo largo de la escala; la segunda transformación es lineal y ubica el punto de corte del nivel de desempeño II en un mismo valor para todos los exámenes: puntuación de 100 en esta escala (cuyo rango va de 60 a 170 puntos²).

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

En la siguiente gráfica puede observarse el número de aciertos obtenidos en dos instrumentos de longitudes diferentes y con puntos de corte distintos que, a partir del escalamiento, es posible graficar en una misma escala, trasladando el primer punto de corte a 100 puntos, aun cuando en cada 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.

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.



4.2 Calificación del resultado global, escala en que se reportará y procedimiento para la conformación de los grupos de desempeño

El resultado global de la evaluación se realizará considerando los siguientes instrumentos:

Para docentes y técnicos docentes:

- o Expediente de evidencias de enseñanza.
- o Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos.
- o Planeación didáctica argumentada.
- o Examen complementario (cuando sea el caso): Segunda lengua: inglés; Segunda lengua: francés; Tecnologías; y Lengua indígena.

Para el personal con funciones de dirección:

- o Expediente de evidencias de la función de dirección
- o Examen de conocimientos y habilidades directivas
- o Ruta de mejora argumentada

Para el personal con funciones de supervisión:

- o Expediente de evidencias de la función de supervisión
- o Examen de conocimientos y habilidades de las funciones de supervisión
- o Plan de trabajo argumentado

El **Informe de cumplimiento de responsabilidades profesionales** tiene por función contar con información sobre el cumplimiento del docente o técnico docente en sus funciones y es proporcionado por el directivo escolar que corresponda. Para fines de la calificación global, este instrumento no será considerado por lo que la información que aporte el directivo escolar será utilizada sólo para efectos de diagnóstico y se dará retroalimentación al docente o técnico docente en el informe individual de resultados de la evaluación, sin afectar su calificación. En el caso de que el docente o técnico docente no tenga el informe, debido a que es responsabilidad de un tercero, se indicará que no se cuenta con la información.

La retroalimentación que recibirá el docente o técnico docente atenderá a los siguientes rubros, según el resultado del análisis del instrumento, a partir de la información disponible: planeación didáctica; atención a alumnos; promoción de ambientes favorables para la sana convivencia; colaboración en la escuela e integración con sus pares; vinculación con padres de familia y normalidad mínima de operación escolar.

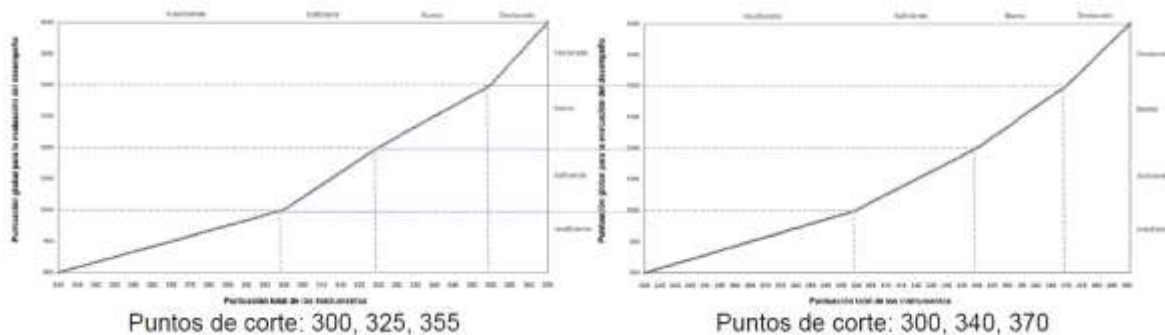
La calificación global

Una vez que se cuenta con las puntuaciones obtenidas en cada uno de los instrumentos de evaluación considerados para la calificación, se procede a calcular la puntuación total de cada sustentante como la suma de dichas puntuaciones. Esto tiene la ventaja de que los aspectos evaluados en que los sustentantes muestran mayores fortalezas compensan aquellos en los que son más débiles, es decir, la calificación está concebida como compensatoria.

Al sumar las puntuaciones obtenidas en cada instrumento de evaluación, los resultados presentan variabilidad, debido a que la escala es aplicable a múltiples instrumentos con características diversas. Por ello, es necesario transformar la suma de las distintas puntuaciones obtenidas a una escala global única de 800 a 1 600 puntos³. Esta operación permite homologar dichas puntuaciones, determinar puntos de corte para establecer los diferentes grupos de desempeño y comunicar de forma sencilla sus calificaciones a los sustentantes.

³ Para mayores detalles sobre el proceso que se lleva a cabo para la transformación de las puntuaciones a la escala global de 800 a 1 600 puntos, consultar el anexo.

A manera de ejemplo, en las siguientes gráficas se muestra la transformación de las puntuaciones en dos trayectos evaluativos diferentes en los que se consideran tres instrumentos para la calificación global; en el eje horizontal de ambas gráficas se muestra que la suma de las puntuaciones de los instrumentos, es decir, la puntuación total, será variable, esto se debe a que depende de la cantidad de instrumentos que considere la evaluación y la puntuación alcanzada en cada uno de ellos. Sin embargo, la transformación que se realiza permitirá comunicar que para el resultado “Suficiente” se requieren, al menos, 1 000 puntos; para el grupo “Bueno”, al menos, 1 200 puntos; y para el grupo “Destacado”, al menos, 1 400 puntos (representados en el eje vertical).



Asimismo, en los ejemplos se observa que en ambas gráficas el punto de corte para alcanzar el resultado “Suficiente” es 300 puntos y la puntuación global está referida a 1 000 puntos. Para alcanzar el resultado global de “Bueno” se requieren 325 puntos en la gráfica de la izquierda y 340 puntos en la gráfica de la derecha, pero ambos están referidos a 1 200 puntos en la escala global. El resultado “Destacado” está referido a 1 400 puntos en la escala global en ambas gráficas, aun cuando en la gráfica de la izquierda se requieren 355 puntos y en la de la derecha 370 puntos.

El resultado “Suficiente”

Criterios para obtener un resultado Suficiente en el proceso de evaluación	
1)	Sustentar todos y cada uno de los instrumentos de la evaluación que son considerados para la calificación
2)	Obtener, al menos, el nivel de desempeño II (N II) en por lo menos dos de los tres instrumentos que integran el proceso de evaluación. En el caso de que la evaluación considere un instrumento complementario, en tres de los cuatro instrumentos; uno de los cuales debe ser éste, es decir: Segunda lengua – inglés; Segunda lengua – francés; o lengua indígena, según sea el caso.
3)	Obtener al menos 1 000 puntos en la escala de calificación global

Todos los instrumentos tendrán la misma jerarquía o peso en la puntuación total de la calificación, por lo que el resultado global de la evaluación del desempeño estará dado considerando la puntuación que aporta cada uno de los instrumentos. De esta forma, el efecto compensatorio de la puntuación global no se ve afectado ni distorsionado por una ponderación diferenciada de los instrumentos de evaluación.

La conformación de los grupos de desempeño

Los grupos de desempeño estarán conformados únicamente por los sustentantes que obtengan, al menos, un resultado “Suficiente” en la evaluación, esto es, hayan sustentado todos y cada uno de los instrumentos de la evaluación que son considerados para la calificación, alcanzado el criterio del número de instrumentos mínimos en N II y, al menos 1 000 puntos en la escala global. Con base en lo anterior, el resultado de la evaluación del desempeño puede ubicarse en alguna de las siguientes categorías:

Resultado de la evaluación: grupos de desempeño

Criterios para formar parte de un grupo de desempeño en Educación Básica	
Resultado de la evaluación	Puntuación global en escala 800 - 1 600
Suficiente	De 1 000 a 1 199 puntos en la puntuación global
Bueno	De 1 200 a 1 399 puntos en la puntuación global
Destacado	Al menos 1 400 puntos en la puntuación global

El resultado “No se presentó a la evaluación”

Para el caso en que el sustentante no presente NINGUNO de los instrumentos considerados para efectos de calificación, 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.

El resultado “Insuficiente”

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

- No sustente alguno o algunos de los instrumentos que son considerados para la calificación.
- **No obtenga**, al menos, N II en dos de los tres instrumentos, o en tres de los cuatro instrumentos cuando sea el caso.
- Haya obtenido **menos** de 1 000 puntos en el resultado global de la evaluación.

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

Se enfatiza que en cualquiera de estos escenarios 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.

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 docente 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, considerando los siguientes términos:

- Serán sujetos a una segunda oportunidad de evaluación del desempeño exclusivamente los Docentes que obtuvieron resultado insuficiente en su primera evaluación del desempeño. Estos docentes presentarán **únicamente** los instrumentos considerados para la calificación en los que haya obtenido nivel de desempeño I en su primera evaluación, así como aquellos a los que no se presentó (si fuera el caso).
- Para garantizar la equidad en los resultados de los docentes sujetos a una segunda oportunidad, se aplicará el mismo esquema de la primera evaluación del desempeño realizada en 2015, es decir, **sólo se considerarán los instrumentos que tuvieron efecto en la calificación** de los docentes en su primera oportunidad de evaluación.
- La calificación global se estimará considerando el resultado obtenido en los instrumentos de evaluación en los que obtuvo, al menos, nivel de desempeño II en su primera oportunidad, y se integrarán con la alcanzada en los instrumentos sustentados en esta segunda oportunidad, siguiendo los mismos principios del modelo de calificación desarrollado en los presentes criterios técnicos (véase la primera sección).
- Para los instrumentos que se sustenten en esta segunda oportunidad, sólo se deben considerar como referencia los puntos de corte empleados en el proceso de evaluación 2016-2017. Lo anterior debido a que los instrumentos de evaluación son sometidos a un proceso de mejora que implica una revisión de los puntos de corte, por lo que previsiblemente éstos tienen ligeros ajustes.

Los Docentes que obtengan un resultado insuficiente en la segunda oportunidad en la evaluación del desempeño, serán reincorporados a los programas de regularización que determine la Autoridad Educativa competente, para sujetarse a una tercera 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.

Anexo técnico

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

Protocolo de calificación por jueces para las rúbricas

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

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

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

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

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

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

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

- a. Se identifica la categoría asignada al sustentante en cada aspecto a evaluar.
- b. Se identifica el valor asignado a cada categoría de la rúbrica.
- c. La suma de los valores es el resultado de la calificación.

9. Las asignaciones de categorías del sustentante en cada aspecto a evaluar para emitir su calificación global 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 multiplicada por uno para el caso de reactivos -toda vez que cada reactivo vale un punto-; o bien, la suma de estas probabilidades multiplicada por el valor máximo posible de las categorías de la rúbrica). Las decisiones de los jueces se promedian obteniendo el punto de corte. La decisión del conjunto de jueces pasa por una primera ronda para valorar sus puntos de vista en plenaria y puede modificarse la decisión hasta llegar a un acuerdo en común.

Método de Beuk

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

Procedimiento

En el cuerpo del documento se señalaron tres fases para el establecimiento de puntos de corte de los niveles de desempeño. Para completar la tercera fase, es necesario recolectar con antelación las respuestas a dos preguntas dirigidas a los integrantes de los distintos comités académicos especializados involucrados en el diseño de las evaluaciones y en 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 de los instrumentos. Adicionalmente, se debe contar con la distribución de los sustentantes para cada posible punto de corte, con la finalidad de hacer converger el juicio de los expertos con la evidencia empírica.

Los pasos a seguir son los siguientes:

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

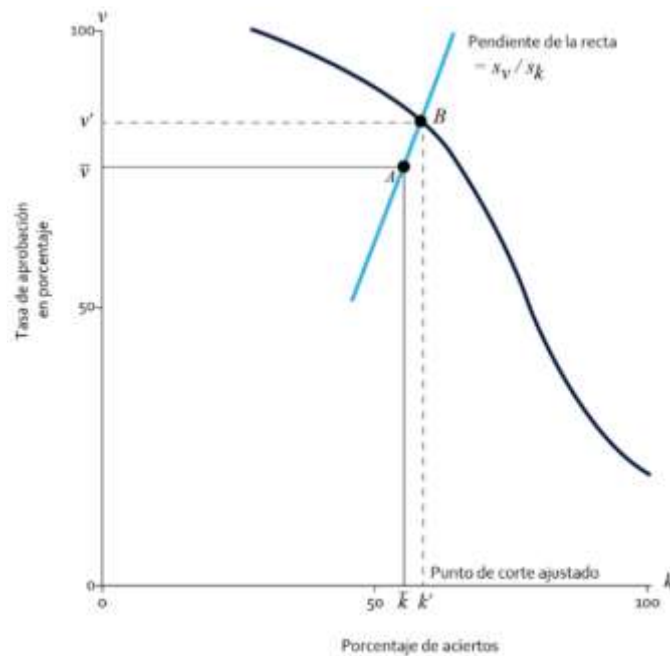
2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

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

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

El punto B, el cual tiene coordenadas (k', v') , representa los valores ya ajustados, por lo que k' corresponderá al punto de corte del estándar de desempeño.

El método asume que el grado en que los expertos están de acuerdo es proporcional a la importancia relativa que los expertos dan a las dos preguntas, de ahí que se utilice una línea recta con pendiente s_v/s_k .



Escalamiento de las puntuaciones

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

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

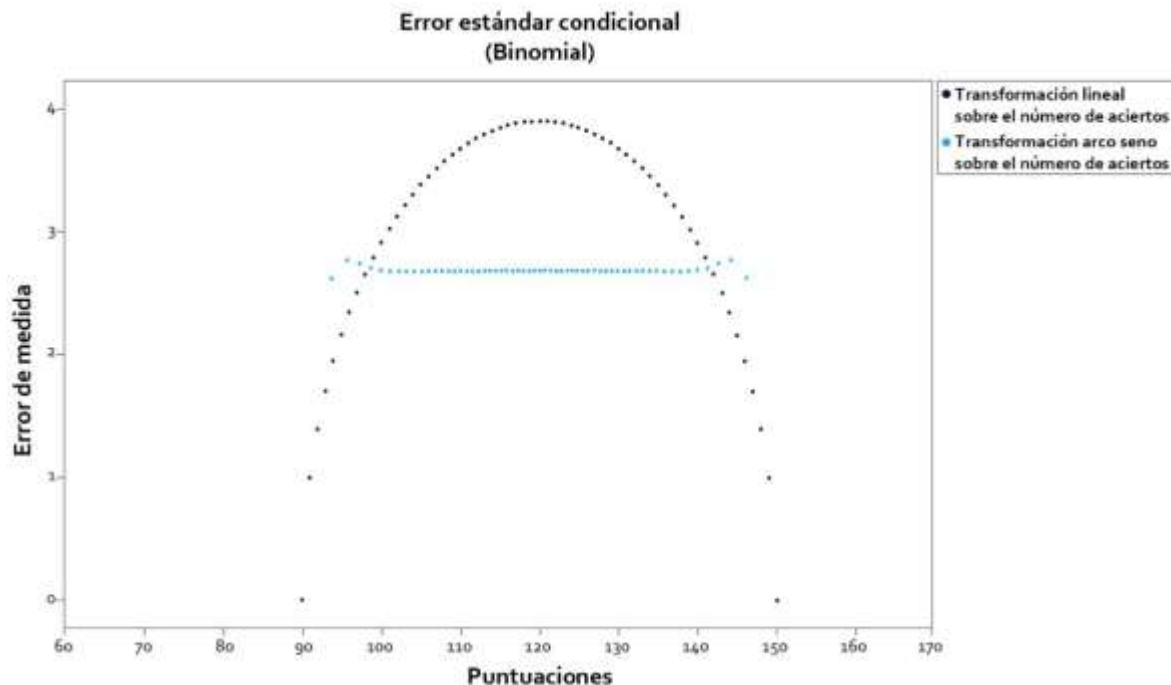
Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta (Kendall y Stuart, 1977), que calcula los errores estándar de medición condicionales, que se describe ulteriormente en este anexo.

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 :

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

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:

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

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:

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

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:

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

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

Procedimiento para el error estándar condicional. Método delta

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

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

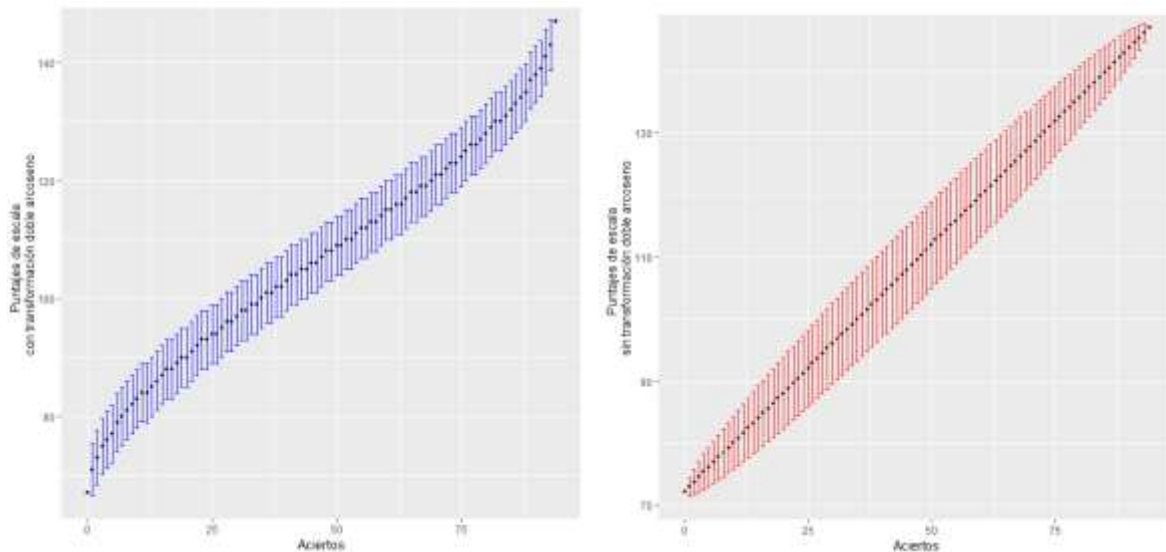
Aplicando lo anterior al doble arcoseno tenemos lo siguiente:

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

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.

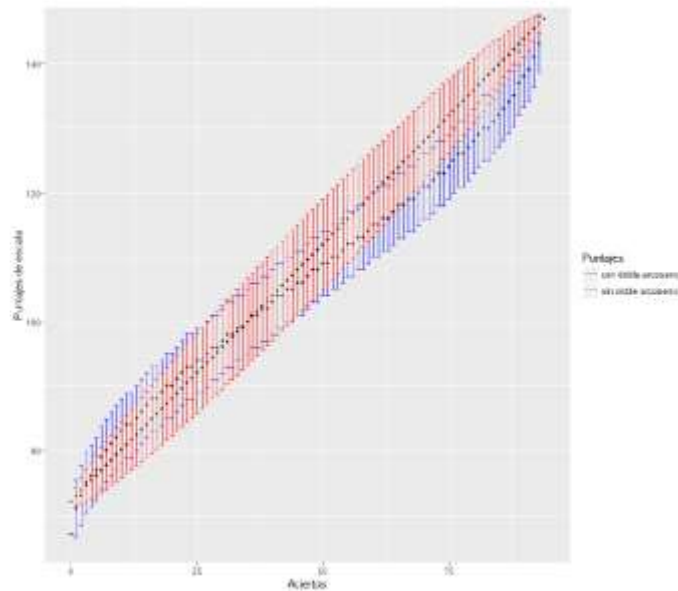
La ventaja de llevar a cabo la transformación doble arcoseno es que el error estándar condicional de medida de los puntajes de la escala se estabiliza y tiene fluctuaciones muy pequeñas, es decir, se mide con similar precisión la mayoría de los puntajes de la escala, a excepción de los extremos. (Brennan, 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, $w_1 + w_2 = 1$ y $w_1, w_2 \geq 0$. Para este proceso se utilizará

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

y

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

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

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

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

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

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

Donde s denota la población sintética y

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

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

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2 \gamma_1^2 [\sigma_1^2(V) - \sigma_2^2(V)] + w_1 w_2 \gamma_1^2 [\mu_1(V) - \mu_2(V)]^2$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1 \gamma_2^2 [\sigma_1^2(V) - \sigma_2^2(V)] + w_1 w_2 \gamma_2^2 [\mu_1(V) - \mu_2(V)]^2$$

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 $x_e = b_1 x + b_0$, 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:

$$\sigma(P_{I_e}) \doteq \frac{A}{2} \left[\frac{b_1}{2(k+1) \left(\sqrt{\frac{x_e}{k+1}} \right) \left(\sqrt{1 - \frac{x_e}{k+1}} \right)} + \frac{b_1}{2(k+1) \left(\sqrt{\frac{x_e+1}{k+1}} \right) \left(\sqrt{1 - \frac{x_e+1}{k+1}} \right)} \right] \sigma(x_e)$$

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).

Algoritmo para el cálculo de la puntuación en escala global

En principio se calcula la puntuación total de los instrumentos, para el siguiente caso se consideran tres para efectos de calificación:

$$PT_i = \sum_{j=1}^{n_i} I_{ji}$$

PT_i = Puntuación total de los instrumentos que alcanza el sustentante i

I_{ji} = Puntuación que alcanza el sustentante i en el instrumento j

$j = 1, 2, 3$ (Instrumento que forma parte de la evaluación [no se considera el informe de responsabilidades profesionales])

$n_i = 3$ (Número total de instrumentos que presenta el sustentante i)

Para cuando se trate de cuatro instrumentos para efectos de calificación, las expresiones matemáticas son como a continuación se describen:

$$PT_i = \sum_{j=1}^{n_i} I_{ji}$$

PT_i = Puntuación total de los instrumentos que alcanza el sustentante i

I_{ji} = Puntuación que alcanza el sustentante i en el instrumento j

$j = 1, 2, 3, 4$ (Instrumento que forma parte de la evaluación [no se considera el informe de responsabilidades profesionales])

$n_i = 4$ (Número total de instrumentos que presenta el sustentante i)

Posteriormente, se establecen los tres puntos de corte globales considerando la escala de puntuaciones PT_i .

Finalmente, se calcula la puntuación en escala global, considerando los puntos de corte establecidos en el paso 2, los cuales serán asociados a 1 000, 1 200 y 1 400 puntos respectivamente, en la escala que va de 800 a 1 600 puntos.

$$\text{Si } \min \{PT_i\} \leq PT_k < PC_1 \quad G_k = 800 + \frac{(PT_k - \min\{PT_i\}) * 200}{PC_1 - \min\{PT_i\}}$$

$$\text{Si } PC_1 \leq PT_k < PC_2 \quad G_k = 1000 + \frac{(PT_k - PC_1) * 200}{PC_2 - PC_1}$$

$$\text{Si } PC_2 \leq PT_k < PC_3 \quad G_k = 1200 + \frac{(PT_k - PC_2) * 200}{PC_3 - PC_2}$$

$$\text{Si } PC_3 \leq PT_k \leq \max \{PT_i\} \quad G_k = 1400 + \frac{(PT_k - PC_3) * 200}{\max \{PT_i\} - PC_3}$$

G_k = Puntuación en escala global del sustentante k en la evaluación del desempeño

PT_k = Puntuación total que alcanza el sustentante k

$\max\{PT_i\}$ = Puntuación total máxima posible

$\min\{PT_i\}$ = Puntuación total mínima posible

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 treinta de septiembre de dos mil dieciséis.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Novena Sesión Ordinaria de dos mil dieciséis, celebrada el treinta de septiembre de dos mil dieciséis. Acuerdo número SOJG/9-16/06, R. La Consejera Presidenta, **Sylvia Irene Schmelkes del Valle**.- Rúbrica.- Los Consejeros: **Eduardo Backhoff Escudero**, **Teresa Bracho González**, **Gilberto Ramón Guevara Niebla**, **Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 439160)