INSTITUTO NACIONAL PARA LA EVALUACION DE LA EDUCACION

CRITERIOS técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados de la evaluación del desempeño del personal docente y técnico docente que ingresó en el ciclo escolar 2014-2015 al término de su segundo año escolar en Educación Básica.

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

CRITERIOS TÉCNICOS Y DE PROCEDIMIENTO PARA EL ANÁLISIS DE LOS INSTRUMENTOS DE EVALUACIÓN, EL PROCESO DE CALIFICACIÓN Y LA EMISIÓN DE RESULTADOS DE LA EVALUACIÓN DEL DESEMPEÑO DEL PERSONAL DOCENTE Y TÉCNICO DOCENTE QUE INGRESÓ EN EL CICLO ESCOLAR 2014-2015 AL TÉRMINO DE SU SEGUNDO AÑO ESCOLAR EN EDUCACIÓN BÁSICA.

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

Así, y con fundamento en lo dispuesto en los artículos 3o. fracción IX de la Constitución Política de los Estados Unidos Mexicanos; 7, fracción 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; Lineamientos para llevar a cabo la evaluación del desempeño del personal docente y técnico docente que ingresó en el ciclo escolar 2014-2015 al término de su segundo año escolar en educación básica y media superior. LINEE-06-2016, la Junta de Gobierno aprueba los siguientes citerios técnicos y de procedimiento para el análisis de los instrumentos de evaluación, el proceso de calificación y la emisión de resultados de la evaluación del desempeño del personal docente y técnico docente que ingresó en el ciclo escolar 2014-2015 al término de su segundo año escolar en Educación Básica (EB).

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 en cuatro apartados: 1) Características generales de los instrumentos para evaluar el desempeño docente; 2) Criterios técnicos para el análisis e integración de los instrumentos de evaluación; 3) Procedimiento para el establecimiento del punto de corte y estándar de desempeño de los instrumentos de evaluación; 4) Resultado de la evaluación del desempeño: resultado por instrumento y resultado global. En la parte final se presenta un Anexo 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 este 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. **Criterios de desempate:** Regla con la cual se determina el orden que ocupan los sustentantes en las listas de prelación, con base en los resultados en los distintos instrumentos que constituyen el proceso de evaluación**.**

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

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

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

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

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

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

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

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

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

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

XVIII. **Especificaciones de tareas evaluativas o reactivos:** Descripción detallada de las características relevantes que se espera 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 estos cuenten con los elementos necesarios para construirlos alineados al objeto de medida o constructo que se desea evaluar a través del instrumento.

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

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

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

XXII. **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.

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

XXIV. **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.

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

XXVI. **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.

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

XXVIII. **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.

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

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

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

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

XXXIII. **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.

XXXIV. **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, evaluado a través de una rúbrica.

XXXV. **Punto de corte:** En instrumentos de evaluación con referencia a un estándar de desempeño, es la puntuación mínima o 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.

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

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

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

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

XL. **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.

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

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

La evaluación del desempeño es un proceso que considera varios instrumentos que permiten identificar la medida en que las personas evaluadas cumplen con los diferentes aspectos que se describen en los Perfiles, parámetros e indicadores. 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 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 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 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 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 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.

Examen complementario

Consiste en un examen que da cuenta de las competencias específicas correspondientes a la función que realiza el docente de secundaria en las tecnologías.

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 del punto 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 de casos 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 solo 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 este proceso de 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 debe ser igual o mayor que 0.80.

  Adicionalmente, para este instrumento, conformado fundamentalmente por escalas tipo Likert, se debe generar evidencia de que los constructos se integran conforme a lo esperado, esto es: a) los reactivos se integran a la o las dimensiones previstas en el diseño del instrumento; b) hay una correlación positiva y significativa entre las distintas escalas que integran el instrumento; c) existe la posibilidad de implementar un modelo de medición a los datos; d) es posible valorar la dimensión del constructo latente y, si es factible e) se verifique que no hay un comportamiento diferencial de los reactivos o las escalas entre subpoblaciones o grupos (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 del punto de corte y estándar 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 este proceso de evaluación, es el establecimiento del punto de corte que divide 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 del punto de corte y estándar 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 establecen los descriptores genéricos de los niveles de desempeño cuya función es orientar a los comités académicos en el trabajo del desarrollo de los descriptores específicos de cada instrumento. Para todos los instrumentos se utilizarán únicamente dos niveles de desempeño: Nivel I (N I) y Nivel II (N II). Los descriptores genéricos para los diferentes grupos de instrumentos y cada nivel se indican en las Tablas 1a, 1b, y 1c.

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 muestra dificultades en el análisis y argumentación de sus evidencias de enseñanza, así como falta de claridad para ajustar su intervención docente en función de las características de los alumnos y de su contexto. Aunque presenta argumentos sobre la elección de los contenidos de aprendizaje por desarrollar en el aula, éstos son poco consistentes con los propósitos educativos, y tiene escasas habilidades para realizar una evaluación pertinente y utilizar sus resultados para mejorar su práctica de enseñanza.
Nivel II (N II)	El docente muestra en el análisis y argumentación de sus evidencias, claridad sobre su práctica de enseñanza y habilidades para ajustar su intervención docente a las características de sus alumnos y de su contexto. Presenta argumentos sobre la elección de los contenidos de aprendizaje a desarrollar en el aula y sobre el impacto que éstos tienen en el aprendizaje de sus alumnos. Demuestra competencias para realizar una evaluación pertinente y utilizar sus resultados para identificar sus fortalezas y áreas de oportunidad, y mejorar de esta manera su práctica de enseñanza.

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 conocimientos poco consistentes acerca de sus alumnos, el currículo, los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana; tiene dificultades para organizar su intervención docente con base en los principios pedagógicos de la Educación Básica. Aunque el docente reconoce algunos conceptos clave del currículo, así como las características generales de la organización de actividades didácticas y el funcionamiento general de una escuela, carece de habilidades para resolver situaciones de su práctica profesional, y para establecer vínculos con la comunidad en la que se encuentra la escuela.
Nivel II (N II)	El docente demuestra conocimientos consistentes acerca de sus alumnos, el currículo, los principios filosóficos, los fundamentos legales y las finalidades de la educación pública mexicana; organiza y sistematiza su intervención docente con base en los principios pedagógicos de la Educación Básica y el reconocimiento de la diversidad cultural de sus alumnos. Resuelve situaciones de su práctica profesional, construye ambientes favorables para la sana convivencia y el aprendizaje de sus alumnos, participa en el funcionamiento eficaz del centro educativo y establece vínculos con la comunidad en la que se encuentra la escuela.

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 muestra dificultades para reconocer las características de sus alumnos, sus intereses y formas de aprendizaje, así como del contexto interno y externo de la escuela. Presenta dificultades para la argumentación y el análisis de la organización y adecuación de los contenidos a desarrollar, y para sustentar su intervención didáctica. Asimismo, muestra conocimientos limitados de las estrategias para evaluar a sus alumnos.
Nivel II (N II)	El docente demuestra conocimientos y habilidades para identificar las características de sus alumnos, sus intereses y formas de aprendizaje, así como del contexto interno y externo de la escuela. Organiza de manera consistente los contenidos a desarrollar, y argumenta y analiza las estrategias didácticas que sustentan su intervención docente. Muestra conocimientos sobre técnicas y métodos de evaluación acordes a los propósitos educativos y emplea sus resultados para mejorar sus prácticas de enseñanza.

Segunda fase

En esta fase se establece el punto de corte a partir del cual se distinguen los niveles de desempeño I y II, en este proceso participan los comités académicos correspondientes a cada instrumento de evaluación. Dichos comités se deberán conformar con especialistas que han participado en el diseño de los instrumentos, garantizando que esté representada 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 reactivos incluidos en las mismas. En todos los casos, el punto de corte se referirá a la ejecución típica o esperable de un sustentante hipotético, con un desempeño mínimamente aceptable para el nivel II. 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 (Angoff, 1971).

Una vez establecido el punto de corte, que divide el rango de calificaciones para diferenciar el nivel I del nivel II 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, es decir, lo que dicho sustentante conoce y es capaz de hacer.

Tercera fase

En la tercera fase se llevará a cabo un ejercicio de retroalimentación a los miembros de los comités académicos con el fin de contrastar sus expectativas sobre el desempeño de la población evaluada, considerando la distribución de sustentantes que se obtiene en cada nivel de desempeño al utilizar el punto de corte definido en la segunda fase, a fin de determinar si es necesario realizar algún ajuste en la decisión tomada con anterioridad (Beuk, 1984).

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

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

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

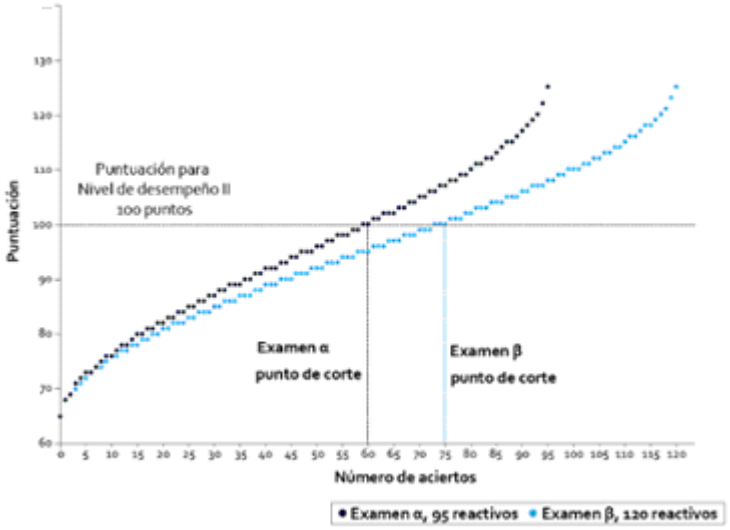
Calificación de los resultados obtenidos por los sustentantes en los distintos instrumentos que constituyen el proceso de evaluación

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

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

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

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



Calificación del resultado global y escala en que se reportará

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

- Expediente de evidencias de enseñanza
- Examen de conocimientos y competencias didácticas que favorecen el aprendizaje de los alumnos de conocimientos disciplinares
- Planeación didáctica argumentada
  - Y, en su caso, el Examen complementario, sólo para el caso en docentes en secundaria en tecnologías

[1] Pueden encontrarse ligeras variaciones en este rango debido a que la escala es aplicable a múltiples instrumentos con características muy diversas, tales como las longitudes, los tipos de instrumentos y su nivel de precisión, diferencias entre los puntos de corte que atienden a las particularidades de los contenidos que se evalúan, entre otras; por otra parte, para realizar el escalamiento, el sustentante debe, al menos, haber alcanzado un acierto en el examen; en caso contrario, se reportará como cero y obtendrá N I. Para mayores detalles sobre los procesos que se llevan a cabo para el escalamiento de las puntuaciones, consultar el anexo técnico.

Como ya se dijo anteriormente, el **Informe de cumplimiento de responsabilidades profesionales** tiene por función contar con información sobre el cumplimiento del 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 en el informe individual de resultados de la evaluación, sin afectar su calificación. En el caso de que el docente no tenga el informe, debido a que es responsabilidad de un tercero, se indicará que no se cuenta con la información.

La calificación global está concebida como compensatoria en tanto que hay un efecto aditivo que permite que las puntuaciones parciales obtenidas en cada uno de los instrumentos utilizados para la evaluación del desempeño, se integren en una puntuación única sobre la que se establecerá el punto de corte global. Este efecto permite que las principales fortalezas de los docentes compensen sus posibles áreas de oportunidad en otros aspectos evaluados. Una vez sumado los aportes que hace cada instrumento que se utiliza para calificar a la puntuación total, se lleva a cabo la transformación a una escala que va de 800 a 1 600 puntos[2].

Todos los instrumentos que son considerados para la calificación tendrán la misma jerarquía o peso en la puntuación total, por lo que el resultado global de la evaluación del desempeño estará dado considerando la puntuación que aporta cada uno de los ellos. 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.

Debido a que en la puntuación global hay sólo un punto de corte, los resultados posibles de la evaluación son los siguientes:

Resultado de la evaluación	**Puntuación global en escala 800 - 1 600**
Insuficiente	Menos de 1 000 puntos en la puntuación global
Cumple con la función docente/técnico docente	1 000 puntos o más en la puntuación global

Para que un sustentante alcance el resultado “Cumple con la función docente/técnico docente”, deberá satisfacer los siguientes criterios:

1) Sustentar todos y cada uno de los instrumentos de la evaluación que son considerados para la calificación.

2) No obtener menos de 100 puntos (nivel NI) en más de uno de los instrumentos considerados para la calificación.

3) Obtener 1 000 puntos o más en la escala de calificación global.

Un sustentante tendrá un resultado “Insuficiente” cuando:

- No presente alguno o algunos de los instrumentos que son considerados para la calificación.
- Obtenga NI en dos o más instrumentos con efecto en la calificación.
- Obtenga menos de 1 000 puntos en el resultado global de la evaluación.

En los dos primeros casos los docentes y técnicos docentes no recibirán información sobre la puntuación global; sin embargo, sí tendrán retroalimentación de cada uno de los instrumentos que hayan sustentado, a fin de que conozcan sus fortalezas y áreas de oportunidad que contribuyan a definir su trayectoria de actualización profesional.

El resultado “No se presentó a la evaluación”

Para el caso en que el 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”.

Consideración final

Es importante destacar que el resultado global de la evaluación es el que debe considerarse como el marco de interpretación para el cumplimiento de la función docente o técnico docente, ya que integra los resultados obtenidos en cada uno de los instrumentos de evaluación.

[2] 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.

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 cada aspecto a evaluar se deben tomar decisiones al respecto, a continuación, se muestran las estrategias que se deben seguir para esta toma de decisiones:

a. Cuando la calificación que se asigna corresponde a categorías de ejecución contiguas (por ejemplo: 1-2) se puede asignar 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 debe asignar 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 y asignarle al sustentante las categorías en cada aspecto a evaluar considerando la decisión del tercer juez y la del juez que había plasmado la decisión más cercana a él. Esto mismo aplica cuando hay reiteradas discrepancias amplias 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 multiplicadas por uno para el caso de reactivos -toda vez que cada reactivo vale un punto-; o bien, la suma de estas probabilidades multiplicadas por el valor máximo posible de las categorías de la rúbrica). Las decisiones de los jueces se promedian obteniendo el punto de corte. La decisión del conjunto de jueces pasa por una primera ronda para valorar sus puntos de vista en plenaria y puede modificarse la decisión hasta llegar a un acuerdo en común.

Método de Beuk

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

Procedimiento

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

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

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

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

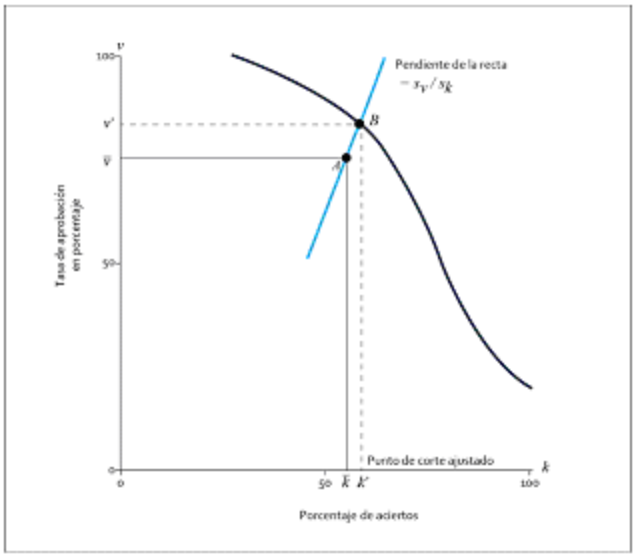
2. Para cada posible punto de corte se grafica la distribución de los resultados obtenidos por los sustentantes en el instrumento de evaluación.

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:

a) Transformación doble arcoseno que estabiliza la magnitud de la precisión que se tiene para cada punto de la escala. **De no implementarla, para cada instrumento se tendría que estimar el error estándar de medida para todas y cada una de las puntuaciones de la escala.**

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

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

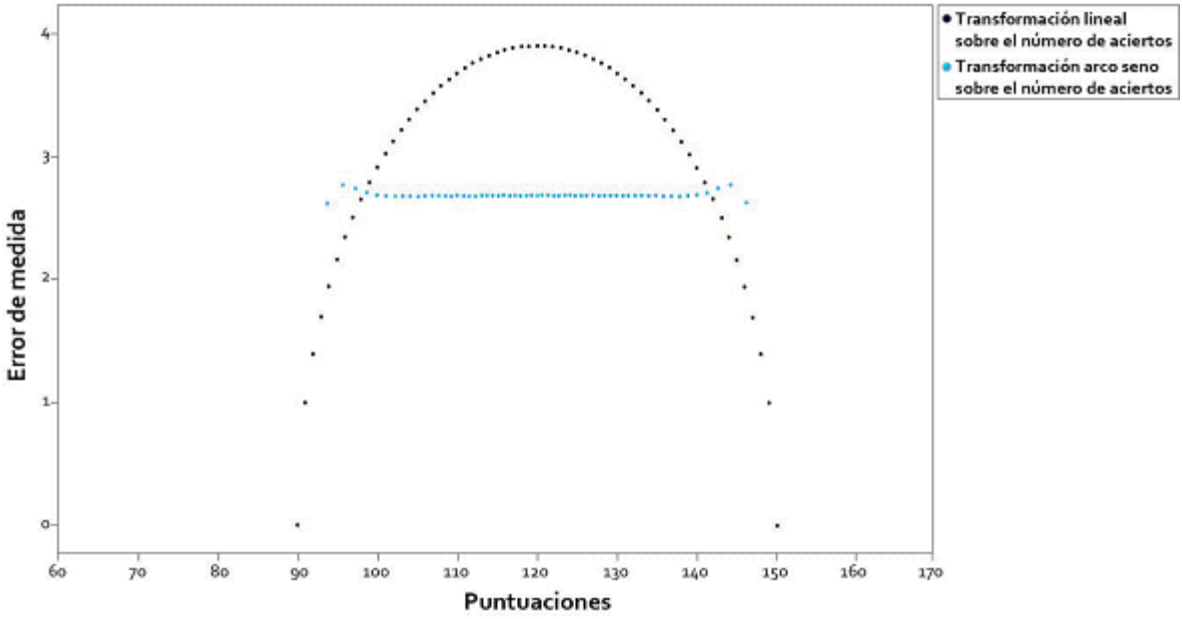
Finalmente, es importante destacar que para que se lleve a cabo el escalamiento, el sustentante debió alcanzar, al menos, un acierto en el instrumento de evaluación en cuestión. De no ser así, se reportará como cero y el resultado será N I.

Procedimiento para la transformación doble arcoseno

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

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

Error estándar condicional (Binomial)



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

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

Donde:

i se refiere a un sustentante

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

K es el número de reactivos del examen

Procedimiento para la transformación lineal

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

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

Donde $\boldsymbol{A = \frac{Q}{[c(K) - c(0)]}}$, $\boldsymbol{B = 100 - A * c(PC)}$, *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 PC es el punto de corte (en número de aciertos) que se definió para establecer los niveles de desempeño y que corresponde al mínimo número de aciertos que debe tener un sustentante para ubicarlo en el nivel de desempeño II.

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

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

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

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

Si el instrumento de evaluación está conformado por dos contenidos específicos, primero se calculará la puntuación del contenido específico 1 (P_{A1i}), mediante la ecuación:

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

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

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

Para los instrumentos de evaluación con más de dos contenidos específicos, se calculará la puntuación de cada uno siguiendo el mismo procedimiento, empleando la ecuación (3) para los primeros. La puntuación del último

contenido específico, se calculará por sustracción como complemento de la puntuación del instrumento de evaluación, el resultado se redondeará al entero positivo más próximo. De esta manera, si el instrumento consta de *j* contenidos específicos, la puntuación del *j-ésimo* contenido específico será:

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

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

Procedimiento para el error estándar condicional. Método delta

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

Para cuantificar el nivel de precisión de las puntuaciones del instrumento, se utilizará el Método delta, (Muñiz, 2003), que calcula los errores estándar de medición condicionales. Para incluir la confiablidad 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.

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, y el error estándar condicional de medición para la transformación $P_{i_e} = A * c(x_e) + B$, que ya incorpora la confiabilidad, está dado por:

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

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

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

La ventaja de llevar a cabo la transformación doble arcoseno es que se estabiliza la magnitud de la precisión que se tiene para cada punto de la escala (Brennan, 2012; American College Testing, 2013; 2014a; 2014b). 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, que establece que los errores estándar de medida condicionales deben reportarse en varios niveles de puntuación, *a menos que haya evidencia de que el error estándar es constante a lo largo de la escala*, lo cual ocurre en este caso, al implementar la transformación doble arcoseno.

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).

Proceso para la equiparación de instrumentos de evaluación

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

Método de equiparación lineal de Levine

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

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

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

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

y

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

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

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

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

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

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

Donde s denota la población sintética y

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

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

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

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

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

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

y

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

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

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

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

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.

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. Como ejemplo se utiliza el caso donde se consideran tres instrumentos para efectos de calificación:

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

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

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

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

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

Posteriormente, se establece el punto de corte global considerando la escala de puntuaciones PT_i.

Finalmente, se calcula la puntuación en escala global, considerando el punto de corte establecido en el paso 2, que será asociado a 1 000 puntos en la escala que va de 800 a 1 600 puntos.

$Si\ min\ \{PT_i\} \leq PT_k < PC$	$G_k = 800 + \frac{(PT_k - min\{PT_i\}) * 200}{PC - min\{PT_i\}}$
$Si\ PC \leq PT_k \leq max\ \{PT_i\}$	$G_k = 1\,000 + \frac{(PT_k - PC) * 600}{max\ \{PT_i\} - PC}$

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

$PT_k = Puntuación\ total\ de\ los\ instrumentos\ que\ alcanza\ el\ sustentante\ k$

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

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

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.

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. 89–107). 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 junio de dos mil dieciséis.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Sexta Sesión Ordinaria de dos mil dieciséis, celebrada el treinta de junio de dos mil dieciséis.- Acuerdo número SOJG/6-16/04, R.- La Consejera Presidenta, **Sylvia Irene Schmelkes del Valle**.- Rúbrica.- Los Consejeros: **Eduardo Backhoff Escudero**, **Teresa Bracho González.**- Rúbricas.

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

(R.- 434492)