INSTITUTO NACIONAL PARA LA EVALUACION DE LA EDUCACION

CRITERIOS técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación.

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

CRITERIOS TÉCNICOS PARA EL DESARROLLO, USO Y MANTENIMIENTO DE INSTRUMENTOS DE EVALUACIÓN

Introducción

El Instituto Nacional para la Evaluación de la Educación (INEE) tiene como tarea principal evaluar la calidad, el desempeño y los resultados del Sistema Educativo Nacional (SEN) en la educación básica y media superior. Para cumplir con ella, debe diseñar y realizar las mediciones de la calidad de los componentes, los procesos y los resultados del SEN, así como expedir los lineamientos a los que se sujetarán las autoridades educativas federales y locales para llevar a cabo las funciones de evaluación que les correspondan. Por esta razón, es preciso que el Instituto cuente con un marco de referencia que permita desarrollar y aplicar instrumentos de evaluación, además de disponer de un referente para valorar la calidad técnica de las evaluaciones que desarrolla o regula. Los presentes CRITERIOS TÉCNICOS PARA EL DESARROLLO, USO Y MANTENIMIENTO DE INSTRUMENTOS DE EVALUACIÓN son una actualización de los publicados en abril de 2014.

Propósito

El propósito de estos Criterios técnicos es proveer referentes para el desarrollo, uso y valoración de la calidad de los instrumentos de evaluación, de las prácticas evaluativas y de los usos de los resultados de las evaluaciones. Aunque la valoración de la calidad de dichos instrumentos depende en gran medida del juicio de profesionales, los Criterios técnicos son un marco de referencia que asegura que los aspectos más importantes de la evaluación educativa sean considerados. En general, estos Criterios buscan proporcionar información técnica relevante de las evaluaciones, de tal manera que los responsables en el desarrollo de los instrumentos, así como las personas involucradas en la toma de decisiones de política educativa conozcan los alcances y limitaciones de los resultados que arrojan los instrumentos de evaluación educativa en el país, y cuenten con una guía para su adecuada interpretación.

Alcance

Estos Criterios técnicos pretenden ser una guía para el desarrollo de cualquier instrumento de evaluación de diferente orden (pruebas, cuestionarios o encuestas) que se elabore dentro o fuera del Instituto y que en su diseño se midan constructos o variables latentes (tales como rendimiento, conocimientos, habilidades -cognitivas, socioemocionales o afectivas-, autoeficacia, representaciones sociales, actitudes, percepciones, entre otros).

El rigor con el que se apliquen los criterios depende de las características de la evaluación; por ejemplo, la exigencia es distinta si se trata de desarrollar instrumentos para una evaluación de rendimiento en la cual se pone en juego la trayectoria académica o profesional de las personas a aquel proceso de evaluación cuyo fin es realizar un diagnóstico de capacidades académicas en el área de ciencias al inicio de un ciclo escolar.

Términos técnicos

A continuación, se enlista el glosario de términos técnicos empleados en el documento para la correcta comprensión de su contenido.

I. **Accesibilidad**: Es el grado en el que las personas con discapacidad son integradas al proceso de evaluación en igualdad de condiciones con las demás.

II. **Adaptaciones al instrumento**: En el contexto de la evaluación, representan los ajustes necesarios al instrumento para garantizar que las personas con alguna discapacidad participen en igualdad de condiciones con el resto de los evaluados.

III. **Administración del instrumento**: Proceso en el que una o más personas contestan el instrumento de evaluación.

IV. **Administrador del instrumento**: Persona responsable de llevar a cabo la aplicación de los instrumentos de evaluación conforme a los protocolos establecidos.

V. **Alto impacto**: Se entiende que una evaluación es de alto impacto cuando sus resultados tienen consecuencias importantes para las personas o las instituciones; por ejemplo, los procesos de admisión o certificación.

VI. **Autoevaluación**: Ejercicio de valoración de las características (conocimientos, actitudes, valores, conducta, etcétera) que la persona que contesta el instrumento hace de sí misma.

VII. **Banco de reactivos**: Repositorio donde se resguardan y clasifican los reactivos que integran los instrumentos de evaluación; en él se administran los datos de identificación del reactivo, sus características métricas, las formas en las que se incorporó y las fechas en las que se utilizó.

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

IX. **Constructo**: Elaboración teórica formulada para explicar un proceso social, psicológico o educativo.

X. **Deseabilidad social**: Se refiere a la tendencia de las personas a dar una imagen más favorable de sí mismas al momento de responder un instrumento de evaluación; lo que ocasiona una distorsión en la medición.

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**: Método estadístico que se utiliza para ajustar las puntuaciones de las formas o versiones de un mismo instrumento, de manera tal que al sustentante le sea indistinto, en términos de la puntuación que se le asigne, responder una forma u otra.

XV. **Error de medida**: Es la diferencia entre el valor medido y el "valor verdadero". Cuando la medida es más precisa, el error es más pequeño y viceversa.

XVI. **Error estándar de medida**: Es la estimación de mediciones repetidas de una misma persona en un mismo instrumento que tienden a distribuirse alrededor de un puntaje verdadero. El puntaje verdadero siempre es desconocido porque ninguna medida puede ser una representación perfecta de un puntaje verdadero.

XVII. **Escala**: Conjunto de números, puntuaciones o medidas que pueden ser asignados a objetos o sucesos con propiedades específicas a partir de reglas definidas.

XVIII. **Escala de actitud**: Conjunto de reactivos que tiene como propósito recolectar información del grado de aceptación o preferencia sobre algún aspecto de interés.

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

XX. **Especificaciones de tareas evaluativas o de reactivos**: Descripción detallada de las tareas específicas susceptibles de medición, que deben realizar las personas que contestan el instrumento de evaluación. Deben estar alineadas al constructo definido en el marco conceptual.

XXI. **Estándar**: Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación. Constituye el referente para emitir un juicio de valor sobre el mérito del objeto evaluado.

XXII. **Estructura**: Está compuesta por los contenidos o aspectos disciplinares que mide un instrumento, así como el número y porcentaje relativo de reactivos o tareas evaluativas con que se integra el instrumento y su distribución.

XXIII. **Evaluación**: Proceso sistemático mediante el cual se recopila y analiza información, cuantitativa o cualitativa, sobre un objeto, sujeto o evento, con el fin de emitir juicios de valor al comparar los resultados con un referente previamente establecido. La información resultante puede ser empleada como insumo para orientar la toma de decisiones.

XXIV. **Formas de un instrumento**: Dos o más versiones de un instrumento que se consideran equivalentes, pues se construyen con los mismos contenidos y especificaciones estadísticas.

XXV. **Funcionamiento diferencial del instrumento (DFT)**: Se refiere a la tendencia del instrumento a funcionar de manera diferente en diferentes subpoblaciones, a pesar de que los individuos que las componen obtengan puntuaciones similares en el instrumento. Las subpoblaciones son definidas por algo distinto a los aspectos relacionados con el constructo evaluado y suelen considerar aspectos de los individuos que las componen, tales como el sexo, la edad, el grupo étnico o el estatus socioeconómico.

XXVI. **Funcionamiento diferencial del reactivo (DIF)**: Se refiere a la tendencia del reactivo a funcionar de manera diferente en diferentes subpoblaciones, a pesar de que los individuos que las componen obtengan puntuaciones similares en el reactivo.

XXVII. **Índice de generalizabilidad**: Este indicador se calcula con el propósito de identificar y estimar la magnitud de las distintas fuentes de variación que pueden intervenir en las diferencias entre puntuaciones o variación debida a las puntuaciones del universo y a las múltiples fuentes de error.

XXVIII. **Instrumento de evaluación**: Herramienta de recolección de datos que suele tener distintos formatos, atendiendo a la naturaleza de la evaluación, por ejemplo, instrumentos de selección de respuesta, instrumentos de respuesta construida, cuestionarios, observaciones, portafolios, entre otros.

XXIX. **Jueceo**: Método en el cual se utiliza la opinión de expertos (denominados jueces) para valorar y calificar distintos aspectos, tales como las respuestas y ejecuciones de las personas que participan en una evaluación o la calidad de los reactivos, las tareas evaluativas y estándares de un instrumento.

XXX. **Mantenimiento**: Conjunto de procedimientos que tienen por objeto conservar actualizado el contenido de un instrumento de evaluación y vigilar su pertinencia, además de nutrir el banco de reactivos y tareas evaluativas con características cualitativas y cuantitativas óptimas.

XXXI. **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.

XXXII. **Modificaciones a las condiciones de aplicación**: En el contexto de la evaluación, representan los ajustes necesarios para garantizar que la administración del instrumento a las personas con alguna discapacidad se lleva a cabo de manera correcta.

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

XXXIV. **Nivel de desempeño**: Criterio conceptual que delimita el marco interpretativo de las puntuaciones obtenidas en una prueba y que refiere a lo que el sustentante es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.

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

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

XXXVII. **Parámetro de referencia**: Indicador estadístico del reactivo o de la tarea evaluativa que se obtiene al considerar un número grande de observaciones que provienen de las administraciones del instrumento ocurridas durante un lapso determinado.

XXXVIII. **Persona que responde el instrumento**: Sujeto que se enfrenta a algún instrumento de evaluación, este concepto incluye a un sustentante que responde una prueba de logro y a un informante en el caso de cuestionarios o encuestas.

XXXIX. **Piloteo de las tareas evaluativas o los reactivos**: Recolección preliminar de datos mediante la administración de un nuevo instrumento de evaluación para valorar su funcionamiento en una muestra de la población objetivo o en una población con características similares a las de la población objetivo, y realizar ajustes orientados a su mejora y a su posterior administración.

XL. **Población objetivo**: Grupo de individuos sobre los cuales se desea que las inferencias elaboradas a partir de los resultados obtenidos con un instrumento de evaluación sean válidas.

XLI. **Protocolo**: Conjunto de reglas o normas que establecen cómo se deben realizar ciertas actividades o pasos, además de incorporar de manera detallada los procedimientos y los estándares que se deben cumplir. Para su realización es fundamental considerar las características especiales de cada evaluación.

XLII. **Prueba**: Instrumento de evaluación que tiene como propósito medir el grado de dominio, conocimiento o aptitud para valorar el mérito de personas, instituciones, programas, sistemas, entre otros.

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

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

XLV. **Sesgo**: Error en la medición de un atributo debido a una variable no controlada, como las diferencias culturales o lingüísticas de las personas evaluadas.

XLVI. **Tabla de especificaciones**: Matriz que permite identificar con toda precisión el objeto de medida o evaluación. Concentra la estructura del instrumento y las definiciones operacionales de lo que se desea medir (especificaciones de reactivos o de tareas evaluativas).

XLVII. **Tarea evaluativa**: Unidad básica de medida de un instrumento de evaluación que consiste en la respuesta que construye una persona o en la ejecución de una actividad, que es susceptible de ser observada y graduada en su nivel de cumplimiento.

XLVIII. **Teoría Clásica de los Tests (TCT)**: Teoría psicométrica que parte del supuesto de que el puntaje observado de una persona que responde un instrumento es la suma de su puntaje verdadero y un error aleatorio independiente del puntaje.

XLIX. **Teoría de la Generalizabilidad (TG)**: Teoría psicométrica que reconoce que existen diferentes fuentes de error de medida y enfatiza la estimación de cada uno por separado; proporciona un mecanismo para la optimización de la confiabilidad denominado coeficiente de generalización, esto es, se centra en los componentes de varianza que indican la magnitud de cada fuente de error que afecta la medición.

L. **Teoría de Respuesta al Ítem (TRI)**: Teoría psicométrica que consiste en una familia de modelos estadísticos que modelan la probabilidad de cierta(s) respuesta(s) (por ejemplo, la respuesta correcta en un test de rendimiento óptimo) como función de las características tanto de la persona evaluada (por ejemplo, su nivel de habilidad en el constructo latente) como del reactivo (por ejemplo, su grado de dificultad).

LI. **Variable latente**: Se denomina así a las variables “ocultas”, es decir, que no son susceptibles de medirse directamente sino a través de otras variables manifiestas (observables).

LII. **Varianza**: La varianza de una variable aleatoria es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

LIII. **Varianza irrelevante para el constructo**: Efecto de variables ajenas al constructo que mide el instrumento de evaluación y que afecta sus resultados; por consiguiente, es información que compromete la validez de las inferencias que se realizan.

Los principios básicos para todo instrumento de evaluación que tiene como propósito la medición son la validez, confiabilidad y equidad. Estos principios aparecen recurrentemente a lo largo del presente documento porque son transversales a todas las fases y pasos del proceso de construcción de cualquier instrumento de evaluación y, en cada una de ellas, se puntualizan los aspectos que se considera fundamental atender y las evidencias que, de manera acumulada, permitan verificar su cumplimiento.

Fases y pasos en el desarrollo de instrumentos de evaluación

Los especialistas en el área de psicometría y evaluación saben que construir un instrumento requiere la aplicación sistemática y minuciosa de ciertos principios que han surgido de la Teoría de la medición. El marco de este documento es el consenso alcanzado entre expertos internacionales en el área de la evaluación psicológica y educativa. Se retoman principalmente dos fuentes. La primera, y más importante, es Standards for Educational and Psychological Testing, desarrollados por la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education. La segunda fuente la constituye el Handbook of Test Development, (Downing y Haladyna, 2006), en el que se proponen una serie de pasos para el desarrollo eficiente de pruebas de rendimiento.

También se consideraron los estándares de calidad técnica propuestos por organismos de evaluación educativa, tales como el Educational Testing Service (ETS); el College Board, la International Test Commission (ITC) y el Joint Committee on Standards for Educational Evaluation (JCSEE), quienes establecen los fundamentos teóricos y las orientaciones que deben tomarse en cuenta para la obtención de evidencias de calidad técnica y validez de las inferencias que se realizan a partir de los resultados de las evaluaciones educativas.

El contenido del documento se organiza en seis fases constituidas por quince pasos, de los cuales se desprenden los criterios técnicos. El procedimiento es secuencial, los productos de una fase se convierten en los insumos de la siguiente, aunque, en la práctica, algunas actividades pueden ocurrir simultáneamente e inclusive, con un orden recursivo.

Es importante destacar que, para el caso de pruebas o bien de los cuestionarios o encuestas, se desprende otra serie de criterios adicionales, que atienden a las particularidades de cada uno de ellos.

Fase 1. Conceptualización del instrumento de evaluación

En esta fase se considera la planeación general y el diseño del instrumento. Cuando se decide desarrollar un instrumento, las interrogantes fundamentales que se deben responder son: ¿qué constructo va a medirse?, ¿cuál será la población objetivo?, ¿quiénes participarán en el desarrollo del instrumento?, ¿qué interpretaciones se realizarán a partir de las puntuaciones obtenidas?, ¿qué formato de instrumento o combinación de formatos será el más apropiado para la evaluación?

El plan general es un marco de referencia sistemático de todas las actividades importantes asociadas con el desarrollo de un instrumento, hace explícitas muchas de las decisiones más significativas, organiza el proyecto en una cronología realista y considera, desde el principio, los problemas relacionados con la seguridad y el control de calidad del instrumento de evaluación.

Paso 1. Planeación general

En este paso se toman decisiones estratégicas respecto a la naturaleza y alcance del instrumento de evaluación que se quiere desarrollar. Se definen en una ficha técnica las características distintivas de la evaluación: propósitos y usos; se determina conceptualmente el constructo o campo de contenido que se pretende medir; se delimita la población objetivo y se establece el modelo psicométrico con el que se analizará el funcionamiento del instrumento (AERA, APA y NCME, 2014).

1.1 **Determinar el propósito de la evaluación.** Es la característica primordial del proceso, responde al para qué se va a evaluar y es el fundamento que da dirección a las actividades que se desarrollarán para construir el instrumento (Schmeiser y Welch, 2006).

1.2 **Definir lo que se pretende medir (objeto de evaluación).** Se establece claramente cuál es el constructo que se desea abordar. Esta definición es la base para el desarrollo del marco conceptual en dónde se fundamenta teóricamente la concepción del constructo que se evaluará.

1.3 **Delimitar la población objetivo en función del propósito.** Se deben indicar las características de las personas a quienes está dirigida la evaluación, para las que son válidas las inferencias que se van a realizar (Mellenbergh, 2011). Cuando se considere necesario, se requiere que también se determinen cuáles son las diferentes subpoblaciones a las que está dirigida la evaluación.

1.4 **Definir la interpretación y uso que se dará a los resultados.** Se debe aclarar qué tipo de inferencias se pueden realizar a partir de los resultados de la evaluación, de tal forma que se delimiten sus alcances.

1.4.1 En el caso de pruebas, si la interpretación se hace de forma normativa (se compara el desempeño de un individuo con el resto de la población que sustentó la evaluación) o con referencia a un criterio (se compara el desempeño del individuo con un estándar previamente establecido).

1.5 **Definir el tipo de instrumento y las modalidades de administración de acuerdo con las características de la evaluación**. A partir del propósito, del constructo a medir, del tipo de inferencias que se desean realizar, los recursos y el tiempo disponible, se debe definir el tipo de instrumento con el cual se recolectará la información: selección de respuesta, respuesta construida o evidencia de desempeño, también se decidirá la manera en la que se administrará: lápiz y papel o dispositivos electrónicos.

1.5.1 En el caso de cuestionarios con escalas de actitud o instrumentos de autoevaluación, se debe prever desde este paso si la medición del constructo podrá contaminarse por deseabilidad social o por el estilo de respuesta de algunas subpoblaciones; esto servirá para planear las estrategias que se implementarán para su control.[1]

[1] De acuerdo con Van de Vijver y He (2014), existen diferentes estilos de respuesta que deben ser controlados: la de aquiescencia (contestar de manera positiva todos los reactivos), intermedia (neutral) y extrema (escoger sistemáticamente las opciones que se encuentran al inicio o al final de la escala).

1.6 **Definir el tipo de respuesta de los reactivos a partir del propósito de la evaluación**. Se debe determinar el formato de respuesta: libre o seleccionada. Si es libre, se deberá determinar lo que se espera que desarrolle la persona que responde el instrumento, así como los criterios de puntuación. Si se opta por respuesta seleccionada, se debe decidir, adicionalmente, el número de opciones de respuesta. Además, es necesario precisar qué se espera que realice el sujeto; por ejemplo, que únicamente marque una opción, que jerarquice las opciones, que seleccione todas las que considere correctas o pertinentes, entre otras; esto tiene implicaciones en el modelo de puntuación que se empleará.

1.6.1 En el caso cuestionarios, se debe también definir los tipos de respuesta y cómo se puntuarán, puesto que en un cuestionario pueden existir diferentes tipos; pueden constituir categorías ordenadas tipo Likert o de elección forzosa o una lista en la cual el orden de las opciones es irrelevante para contestar.

1.7 **Establecer el modelo de medición en función del alcance de la evaluación**. Se debe seleccionar el modelo de medición mediante el cual se realizará el análisis de reactivos y los aspectos relacionados con la confiabilidad, la precisión, la validez de las inferencias y la detección de posibles sesgos de la evaluación. Es importante considerar el tamaño de la muestra a la que se aplicará el instrumento.

1.8 **Definir las especificaciones métricas de los parámetros e indicadores de los reactivos o tareas y del instrumento**. En función del modelo de medición seleccionado se establecen las propiedades estadísticas deseadas para los reactivos o tareas, así como las del instrumento en su totalidad.

1.9 **Definir la longitud del instrumento en función de la evaluación y de las condiciones en las que ésta se desarrollará**. Se debe determinar el número de reactivos o de tareas evaluativas que integrarán el instrumento, para ello se deben tener en cuenta las inferencias que se harán con los resultados, el uso de los mismos y el error de medida que se desea asumir, además de las restricciones logísticas y temporales para la administración (Schmeiser y Welch, 2006).

Considerar que no necesariamente un instrumento con un gran número de reactivos o tareas evaluativas es el mejor. Aunque una manera de mejorar la confiabilidad de un instrumento es incrementar el número de reactivos o tareas evaluativas, de acuerdo con Thompson (1990) y Embreston y Reise (2000), pueden existir instrumentos cortos con mejores índices en su métrica que uno extenso, todo depende de la calidad técnica de las unidades de medida (reactivos o tareas) que los integren.

1.10 **Definir la organización del instrumento de acuerdo con su contenido, longitud y tiempo de administración**. Determinar con antelación la manera en la que se hará el ensamble de los cuadernillos y las formas, así como su empaquetamiento, con base en la logística de administración del mismo. Por ejemplo, si el instrumento se aplicará en dos sesiones, debe preverse su división en dos cuadernillos independientes (Schmeiser y Welch, 2006).

1.10.1 En el caso de pruebas, determinar el número de formas a ensamblar con base en el alcance de la evaluación. Se debe decidir para cada periodo de aplicación si se desarrollará una única forma del instrumento, o bien, si se construirán formas alternas equivalentes, en cuyo caso, se debe especificar el número de éstas, indicando si serán cuadernillos distintos u ordenamientos.

1.11 **Integrar los cuerpos colegiados.** Los cuerpos colegiados constituyen los grupos de expertos que participarán realizando diversas tareas en las distintas fases de construcción del instrumento. El principio que debe guiar su conformación es el de contar con especialistas del campo o dominio a evaluar y algunos que atiendan a otros aspectos relevantes para la validez del instrumento (sesgo, atención a la diversidad, diseño universal[2]). La formación y actividad profesional de estos especialistas debe garantizar el desarrollo de un instrumento cuyos resultados permitan hacer inferencias válidas.

En la conformación de los cuerpos colegiados deben estar representadas las subpoblaciones o subsistemas a los que va dirigida la evaluación (región geográfica, grupo étnico, etcétera) y debe tenerse en cuenta la función específica que cada uno de ellos cumplirá en las distintas fases del proceso de desarrollo del instrumento. Los integrantes de estos cuerpos colegiados deben estar capacitados para realizar las tareas específicas que les sean asignadas durante el desarrollo del instrumento de evaluación.

[2] Es el diseño de productos, entornos, programas y servicios para ser usados por el mayor número de personas sin necesidad de adaptación ni diseño especializado. No excluye la ayuda técnica para grupos particulares de personas con discapacidad cuando así se requiera (Ley General para la Inclusión de las Personas con Discapacidad, [DOF, 2011]).

Las denominaciones propuestas son sólo sugerencias que intentan expresar la naturaleza de la función y ámbito de responsabilidad de cada cuerpo colegiado, pero pueden adoptar distintas formas en cada caso. Por ejemplo, lo que aquí se propone como consejo rector, puede adquirir la denominación de Consejo Técnico o Consejo de Supervisión, lo importante es entender que debe existir un órgano colegiado que se encargue de estas funciones.

Por otra parte, es posible que, dada la complejidad y factibilidad que representa la organización de cada uno de estos cuerpos colegiados, algunos especialistas puedan participar en el Consejo Rector del Instrumento y en más de un comité de apoyo, con la finalidad de dar coherencia y continuidad al trabajo y a los productos que se generen en cada fase.

Sin embargo, de estos cuerpos colegiados, aquellos que son indispensables para cualquier tipo de instrumentos y que se constituirían si hay los recursos necesarios (tiempo, financieros, humanos) para atender aspectos particulares de la evaluación, es indispensable que, al menos, cuatro de ellos sean constituidos: Consejo Rector del Instrumento (que puede contribuir al diseño del instrumento), el Comité de especificaciones (que operacionaliza el objeto de medida y dicta las pautas para la construcción de los reactivos y tareas evaluativas), así como los comités de elaboración y validación de reactivos (quienes de manera plural pueden incorporar elementos de cotejo para la revisión de que los reactivos o tareas evaluativas estén libres de sesgo y sean adecuados para la población objetivo).

A continuación, se señalan las funciones que cada grupo debe desempeñar:

a) Consejo Rector del Instrumento. Es el cuerpo colegiado de mayor jerarquía, avala las características principales de la evaluación, tales como el marco conceptual del instrumento, el objeto de medida y el establecimiento de los puntos de corte, si es el caso. Asimismo, vigila que las actividades de los comités estén alineadas a los aspectos centrales del instrumento. Es el encargado de aprobar la ficha técnica y de avalar el marco teórico del instrumento de evaluación, así como de proponer expertos que integren los distintos cuerpos colegiados o comités académicos (en sus distintas denominaciones), para desarrollar las diferentes tareas sustantivas en el desarrollo de la evaluación.

b) Comité de diseño. Su función es participar en la construcción del marco conceptual, además de seleccionar, delimitar conceptualmente y justificar el objeto de medida o contenido del instrumento de evaluación.

c) Comité de especificaciones. Tiene como función precisar y operacionalizar el objeto de medida del instrumento de evaluación, a partir de la redacción de especificaciones.

d) Comité de elaboración de reactivos o tareas evaluativas. Su propósito es elaborar los reactivos o las tareas evaluativas a partir de las especificaciones previamente desarrolladas.

e) Comité de validación. Tiene como función verificar que los reactivos o las tareas evaluativas del instrumento estén debidamente alineados con las especificaciones, que no presenten errores de contenido y que se redacten en un lenguaje apropiado para la población evaluada.

f) Comité de adaptaciones y modificaciones. Su función es asesorar al organismo evaluador en los casos en que se realicen adecuaciones al instrumento o a las condiciones de su administración, para atender a poblaciones con discapacidad[3] o hablantes de una lengua distinta al español.

g) Comité de sesgo. Su función consiste en vigilar que, durante el desarrollo de la evaluación, el instrumento, los materiales dirigidos a la persona que responde el instrumento y los materiales dirigidos a los usuarios de la evaluación no presenten información ofensiva, así como que su contenido no favorezca a algún sector de la población y esté libre de estereotipos de género o culturales. Su trabajo debe permear a lo largo de las fases y pasos que aquí se describen y el organismo evaluador determinará específicamente los momentos en los que este cuerpo colegiado participará, así como las evidencias que sustentarán su participación de manera transversal.

h) Para el caso de pruebas, se constituye de manera particular para los instrumentos de respuesta construida el Comité de asignación de puntuaciones de las tareas evaluativas. Cuando la puntuación descansa en los juicios de evaluadores, la función de este comité es asignar de manera objetiva una puntuación al sustentante con base en los criterios establecidos exprofeso; los miembros del comité deben poseer los conocimientos y la experiencia requeridos y determinan la categoría o codificación que debe asignarse con base en los distintos niveles de ejecución que se evalúan a través de las tareas.

[3] Persona con discapacidad: Individuo que por razón congénita o adquirida presenta una o más deficiencias de carácter físico, mental, intelectual o sensorial, ya sea permanente o temporal y que al interactuar con las barreras que le impone el entorno social, pueda impedir su inclusión plena y efectiva, en igualdad de condiciones con los demás (Ley General para la Inclusión de las Personas con Discapacidad [DOF, 2011]).

1.12 **Determinar las estrategias para atender a poblaciones con necesidades especiales**. Para establecer cuáles son las adecuaciones necesarias en la evaluación es indispensable analizar el propósito, la población objetivo y las inferencias que se desea realizar con los resultados; se debe identificar cuáles necesidades especiales[4] se quiere cubrir (considerar discapacidades que no estén relacionadas con el objeto de medida del instrumento), el costo que implicará el ajuste y el impacto que éste tendrá, en función del número de evaluados que serán atendidos (estudio de factibilidad). Esto debe realizarse desde la planeación general del instrumento y, una vez fundamentada la necesidad de hacer una adaptación, considerarla en los pasos subsecuentes (ETS, 2015).

1.13 **Ofrecer adaptaciones a las formas del instrumento apropiadas para las personas con discapacidad**. Se debe justificar de manera formal la estrategia para realizar las formas adaptadas del instrumento, con base en una investigación sobre las necesidades individuales, no en una elección al azar (Downing y Haladyna, 2006). También se debe determinar qué adecuaciones compensarán el sesgo ocasionado por la discapacidad sin proporcionarles una ventaja injustificada respecto a lo que se pretende evaluar (Koretz, 2010).

1.14 **Determinar las adaptaciones del instrumento a personas que hablan otra lengua.** El instrumento debe administrarse en la lengua más relevante y adecuada a la luz de los objetivos de la evaluación. Si la población objetivo incluye a personas que crecieron dentro de una cultura en la que domina otra lengua, de ser factible, y en la medida de lo posible, se deben proveer formas alternas del instrumento para estas lenguas o recurrir a intérpretes.

1.15 **Definir el procedimiento para determinar la puntuación del instrumento.** Se establece la manera en la que se puntuará el instrumento (por ejemplo, a través de un modelo de medición dicotómico o bien, politómico). También se debe elegir la manera en que se van a interpretar las puntuaciones (por ejemplo, de manera criterial,[5] normativa[6] o ipsativa[7]).

1.15.1 En el caso de pruebas, se debe definir el modelo de calificación global del instrumento y el tipo de escalamiento que se implementará (Holland y Strawderman, 2011): si el modelo de calificación es analítico es necesario definir cómo se combinan las calificaciones parciales; es decir, si se aplican ciertas ponderaciones, o bien, si cada calificación parcial tiene el mismo peso para la global. Si es compensatorio, una calificación baja en una dimensión se contrarresta con una calificación alta en otra o es conjuntiva cuando se consideran con el mismo valor todas las calificaciones parciales y se unen para determinar la calificación global (Council of Europe, 2011).

1.16 **Elaborar el marco teórico o conceptual de la evaluación.** Este documento tiene como propósito principal la definición teórica o conceptual del constructo; a partir del propósito de la evaluación y del tipo de inferencias que se desea realizar, se debe adoptar una conceptualización o explicación teórica del rasgo que se va a medir y justificar por qué esa definición es la más adecuada en el contexto de la evaluación. Esta visión se sustentará en una revisión lo más amplia posible de la literatura y se desagregarán los dominios que, de acuerdo con esa postura, integran el constructo. Este documento es fundamental para sustentar el objeto de medida que se formaliza en la estructura y en las especificaciones de los reactivos o de las tareas evaluativas (AERA, APA y NCME, 2014; Mellenbergh, 2011).

1.17 **Revisar la sensibilidad del instrumento y conservar las evidencias de este trabajo.** En la planeación es necesario verificar el contenido y el lenguaje utilizado en el instrumento y asegurarse que sea apropiado y respetuoso, así como esté libre de sesgo cultural y lingüístico. Este escrutinio es independiente a la verificación del contenido que se hace en la validación de reactivos y de los aspectos técnicos; deben participar tanto hombres como mujeres e incorporar representantes de los grupos minoritarios, tales como los grupos étnicos y lingüísticos de los cuales se desea obtener información (Mellenbergh, 2011).

---

[4] Considerar todos los tipos de discapacidad: caminar, subir o bajar usando sus piernas; ver (aunque usen lentes); mover o usar sus brazos o manos; aprender, recordar o concentrarse; escuchar (aunque usen aparato auditivo); bañarse, vestirse o comer; hablar o comunicarse y problemas emocionales y mentales (INEGI, 2015).

[5] Criterial, se compara a la persona que responde la prueba con un estándar previamente establecido.

[6] Normativa, se compara al sujeto con el resto de las personas evaluadas.

[7] Ipsativa, se compara al sujeto contra sí mismo; este tipo de puntuación es utilizada en mayor medida en cuestionarios de autoevaluación.

1.18 **Garantizar la idoneidad del personal que coordina y participa en la construcción del instrumento.** Se debe asegurar que el responsable de dirigir el desarrollo y construcción del instrumento cuente con los conocimientos académicos y la experiencia profesional necesarios para cumplir con esta función. Asimismo, demostrar que las personas, que participan de manera activa en el desarrollo de alguna actividad vinculada con la construcción del instrumento, poseen los conocimientos y las habilidades requeridas para desempeñar adecuadamente su función como parte de alguno de los cuerpos colegiados indicados previamente.

1.19 **Dar a conocer al personal involucrado en el desarrollo de la evaluación las políticas de seguridad**. Es responsabilidad de la instancia técnica que coordina la construcción del instrumento de evaluación, informar, a todas las personas que de alguna manera colaboren en el desarrollo de la evaluación, las medidas para guardar la confidencialidad del instrumento y de la información derivada de éste, así como las estrategias de auditoría que se implementarán.

Evidencias para la verificación documental

- ✓ **Ficha técnica del instrumento de evaluación firmada por el Consejo Rector del Instrumento**. Debe contener los siguientes aspectos: definición de lo que se pretende medir; propósito y población objetivo; uso y alcance de sus resultados; tipo de instrumento, impacto o consecuencias de sus resultados; usuarios de la información; características específicas del instrumento como formato, modalidad y condiciones de administración; longitud del instrumento y racionalidad bajo la cual se determinó; tratamiento que se le dará a las respuestas; marco teórico de la medición; puntuación del instrumento; modificaciones a las condiciones de la administración y adaptaciones del instrumento. También se deben incorporar los perfiles de los comités académicos que participarán en el desarrollo de la evaluación.
- ✓ **Marco teórico o conceptual del instrumento firmado por el Consejo Rector del Instrumento**. Este documento debe delimitar de manera teórica o conceptual, el constructo que se abordará en el instrumento, además de referir las dimensiones que lo componen, su definición y la relación de cada una de ellas con el constructo. También se debe explicar cómo se abordará el constructo a través del instrumento seleccionado, así como la justificación de esta selección.
- ✓ **Acta de instauración del Consejo Rector del Instrumento y minutas que documentan sus sesiones de trabajo.** Deben recuperar los acuerdos tomados en la sesión respectiva y estar firmadas por los miembros del cuerpo colegiado.
- ✓ **Currículum Vítae de los integrantes del Consejo Rector del Instrumento**. Debe sustentar que las personas que componen este cuerpo colegiado son expertas en el contenido del instrumento o en algún aspecto relacionado con su construcción.

Evidencias adicionales en el caso de cuestionarios o instrumentos con escalas

- ✓ Documento donde se describa el (los) tipo(s) de respuesta utilizado(s) y, si es el caso, la estrategia que se implementará para controlar la deseabilidad social o para detectar patrones o estilos de respuesta de alguna subpoblación que pudieran influir en los resultados de la evaluación.
- ✓ Si se utiliza un método de puntuación ipsativa, presentar un documento que justifique las fuentes de la medición con las que se realizará la comparación.

Paso 2. Diseño del instrumento

En este paso se operacionaliza el constructo que se pretende medir, se determina el contenido del instrumento al establecer la estructura y las especificaciones de los reactivos o de las tareas evaluativas, lo que en su conjunto constituye la tabla de especificaciones. En el diseño de la evaluación se toma como referencia el contenido del marco conceptual o teórico del instrumento elaborado en el paso anterior, así como la ficha técnica y, en términos observables, se define qué se va a medir y cómo. A partir del objeto de medición del instrumento, se elaboran y validan los reactivos o las tareas evaluativas. En los casos de instrumentos de respuesta construida, es en este paso donde se define la rúbrica y los niveles de ejecución que se considerarán en ésta.

2.1 **Documentar el procedimiento que se seguirá para la conformación de la tabla de especificaciones.** Se debe establecer de manera formal y fundamentada el método que se seguirá para desdoblar el objeto de medida del instrumento en una tabla de especificaciones. Es decir, el procedimiento, con base en la estructura, para la distribución de reactivos o de tareas evaluativas y su justificación. Posteriormente se deberá incluir el método para derivar las especificaciones de reactivos o tareas.

2.2 **Capacitación de los comités académicos de diseño**. Una vez instaurado el cuerpo colegiado que se encargará de la delimitación conceptual del objeto de medida, así como de su operacionalización a través de las especificaciones, se le darán a conocer los insumos y los lineamientos técnicos necesarios para desarrollar adecuadamente sus funciones. Esta capacitación incluye la evaluación de los especialistas para asegurar la plena comprensión de lo que se espera de ellos y de los productos de su trabajo colegiado.

2.3 **Definir los elementos de la estructura del instrumento.** El contenido de la prueba debe consignarse en una matriz en la que se ordenan, de lo general a lo particular, las categorías temáticas a medir. Para cada uno de los niveles considerados en la estructura, se debe establecer el número de tareas evaluativas o de reactivos que se integrarán en el instrumento, respetando la longitud determinada en el paso anterior. Se debe evidenciar la secuencia de las secciones y de los reactivos. Posteriormente, se debe definir la proporción de reactivos correspondientes a cada nivel de la estructura (ponderaciones), lo cual debe reflejar la importancia relativa de cada uno de los contenidos en el conjunto del instrumento.

2.3.1 En el caso de pruebas, la estructura debe tener, al menos, tres niveles de desagregación para instrumentos de opción múltiple (por ejemplo, áreas, subáreas y temas o aspectos por evaluar); el primero y el segundo nivel deben contar, como mínimo, con dos conjuntos de contenidos específicos; por ejemplo, para cada área dos subáreas y para cada subárea dos temas y se debe definir la proporción de reactivos en cada nivel de desagregación, ya que esto permite construir formas del mismo instrumento a partir de los mismos contenidos, longitud y ponderación, lo que las hace cualitativamente equivalentes. En el caso de instrumentos de respuesta construida, se deben considerar al menos dos niveles de desagregación; por ejemplo: áreas y subáreas.

2.4 **Elaborar las especificaciones de reactivos o tareas evaluativas.** Las especificaciones constituyen la operacionalización de lo que se va evaluar, proporcionan los elementos necesarios para interpretar la profundidad y el alcance de los contenidos y determinar el formato de reactivo que permita atenderla a cabalidad (Haladyna y Rodríguez, 2013). Éstas juegan un papel protagónico en la elaboración y validación de los reactivos que conformarán el instrumento, puesto que son el referente directo entre el reactivo o la tarea evaluativa y el constructo. Las especificaciones deben incluir, al menos: a) la definición operacional (de la cual puede identificarse el conocimiento o habilidad que se busca evaluar, el nivel de complejidad o proceso cognitivo, así como la situación contextual que le dé precisión); b) un ejemplo del reactivo o tarea evaluativa con la fundamentación de la respuesta óptima o correcta, así como de los distractores o respuestas parciales y c) la bibliografía de apoyo.

2.5 **Integrar la tabla de especificaciones y vigilar que el constructo esté adecuadamente representado.** A partir de la distribución definida en la tabla de especificaciones, se debe asignar el número de reactivos o tareas evaluativas con las que se va a valorar cada aspecto a evaluar. Una vez elaboradas las especificaciones, se deben valorar de manera integrada para verificar que cumplen con los elementos requeridos, haya consistencia interna y no exista redundancia en lo que se busca medir. Se debe tener cuidado de incorporar lo realmente importante para el objeto de medida en cuestión, puesto que en la evaluación se incluye sólo una muestra de los indicadores que lo componen (Koretz, 2010).

2.5.1 En el caso de los instrumentos de respuesta construida, se debe establecer la rúbrica o guía de calificación a partir de los siguientes componentes:

- Aspectos por evaluar (los rasgos que van a ser valorados).
- Niveles de habilidad, ejecución o competencia (indicando la puntuación asignada a cada aspecto por evaluar y que representa adecuadamente el juicio acerca de qué tan buena fue la ejecución de la tarea).

Aspectos por evaluar	Nivel de habilidad, ejecución o competencia
	¿Cómo decidir si un aspirante hace bien o no la tarea?

2.5.2 En las rúbricas o guías de calificación, los distintos niveles o categorías de ejecución que se consignen deben ser claramente distinguibles entre sí y con un diseño ordinal ascendente (de menor a mayor valor) o descendente (de mayor a menor valor), dependiendo del constructo que se busca medir.

2.5.3 En la construcción de una rúbrica se debe:

- Describir en términos cualitativos los niveles de habilidad, ejecución o competencia. Las descripciones de cada nivel deben ser claras y diferenciables, deben proporcionar los suficientes elementos para entender lo que se espera observar típicamente en ese grado de dominio. No es adecuado sólo adicionar o eliminar elementos de los descriptores, a medida que el dominio aumenta o disminuye. Al leer todas las etiquetas debe apreciarse una progresión clara del dominio. Si es posible, no utilizar más de cuatro niveles para facilitar los procesos de calificación (Perie, 2008).
- Describir la tarea. Dar una visión general en dos o tres enunciados de lo que los evaluados deben hacer.
- Escribir las instrucciones para el sustentante. Las consignas o tareas deben contemplar, de forma clara y concisa, todos los elementos que serán considerados en la rúbrica de calificación. Precisar la longitud y el alcance del producto requerido. Desarrollar los aspectos por evaluar. Elaborar una lista de elementos del producto que se espera, en orden secuencial o por grado de importancia; deben establecerse antes de desarrollar la rúbrica y vincularse directamente con los componentes de la tarea a desarrollar. Basar los aspectos por evaluar en una investigación previa, estándares, percepciones de los involucrados, etcétera. Utilizar el mismo lenguaje que se usó en las instrucciones.
- Establecer las puntuaciones para cada nivel de ejecución. Se deben establecer las puntuaciones para cada nivel de habilidad, ejecución o competencia y describir la manera en que se determina la puntuación total de los evaluados. Incluir orientaciones o ejemplos de lo que se espera que desarrolle el sustentante en cada nivel de ejecución.

2.6 **Documentar imponderables durante el proceso de desarrollo del instrumento.** Cualquier decisión no prevista que se tome durante el desarrollo del instrumento, debe estar justificada con evidencias teóricas o empíricas.[8] Las restricciones logísticas y pragmáticas pueden formar parte de la justificación.

2.7 **Realizar la revisión y corrección de estilo de la estructura y las especificaciones.** Antes de someter estos documentos a la aprobación del Consejo Rector del Instrumento, se debe revisar que no haya errores ortográficos ni de redacción. Se deberá seguir una línea editorial definida con el propósito de homologar el estilo de la redacción y garantizar su claridad (INEE, junio 2016).

2.8 **Documentar el trabajo y las decisiones de los comités académicos**. Las decisiones tomadas por los cuerpos colegiados en las distintas sesiones de trabajo deben ser registradas en bitácoras, actas o minutas y éstas deben ser firmadas por todos sus integrantes.

[8] La evidencia teórica puede ser una referencia al marco teórico o conceptual del instrumento u otros modelos en la literatura especializada. La evidencia empírica puede ser de experiencias previas con evaluaciones similares, por ejemplo, las especificaciones métricas más apropiadas; también estudios empíricos en la literatura pueden guiar y justificar el uso de ciertos formatos de reactivos o respuestas, el número de opciones en formatos de respuesta seleccionada.

Evidencias para la verificación documental

- ✓ **Procedimiento empleado para delimitar el objeto de medida**. Este documento debe enunciar la estrategia para definir qué se va a medir y cómo, incluir los fundamentos para realizar este procedimiento en específico e incorporar la manera en la que se realizará la ponderación de cada nivel de la estructura para la distribución de los reactivos.
- ✓ **Material de capacitación de los cuerpos colegiados.** Como evidencia de que se realizó la inducción de los comités que participan en este paso, se debe resguardar la carta descriptiva de las sesiones de capacitación, en la cual se incluyan los temas revisados y el tiempo dedicado a cada uno de ellos, así como los ejercicios realizados para favorecer su comprensión.
- ✓ **Tabla de especificaciones firmada por el comité correspondiente**. Debe concentrar la estructura y las especificaciones de reactivos o de tareas evaluativas, establecer los contenidos organizados en niveles de desagregación, ser congruente con el procedimiento determinado para su delimitación e incluir la distribución de reactivos hasta el último nivel. Las especificaciones deben ser congruentes con el tipo de instrumento que se desarrollará.
- ✓ **Bitácoras que documentan el trabajo de los comités académicos**. Las sesiones de trabajo deben ser respaldadas con listas de asistencia y documentos que describan las actividades desarrolladas en cada reunión; los acuerdos tomados deben ser firmados por los miembros del cuerpo colegiado.
- ✓ **Currículum Vítae de los integrantes de los comités académicos de diseño y de elaboración de especificaciones**. Este documento sustenta que las personas que participaron en este paso son expertas en el contenido del instrumento y tienen una trayectoria profesional que les permite cumplir las funciones establecidas.

Evidencias adicionales para el caso de pruebas

- ✓ **En los instrumentos de respuesta construida, deberán estar firmados por el Consejo Rector del Instrumento.** Está compuesto por niveles de ejecución –no más de cuatro–, por la puntuación establecida para cada nivel de ejecución, por los descriptores de desempeño y los aspectos a evaluar. Además de las consignas o tareas evaluativas proporcionadas a las personas que sustentarán la prueba.
- ✓ **En los instrumentos de respuesta construida, el documento que sustenta la complejidad de las especificaciones.** Debe describirse la taxonomía cognitiva a partir de la cual se redactaron las especificaciones, esta clasificación explicará los diferentes niveles de dominio que se desea explorar en el instrumento. Cada nivel debe ser excluyente de otro, de modo que sea clara la diferencia de las acciones que le corresponden.

Fase 2. Desarrollo del instrumento de evaluación

En esta fase se abordan las actividades involucradas en la elaboración y piloteo de las tareas evaluativas o de los reactivos y se integran los criterios relacionados con el ensamble del instrumento. Durante esta fase también se realizan diferentes revisiones técnicas y de contenido para garantizar que los reactivos y las tareas evaluativas no tengan errores y estén libres de varianza irrelevante al constructo, además de verificar que corresponden con el objeto de medida.

Paso 3. Elaboración de las tareas evaluativas o de los reactivos

Se describen las actividades necesarias para constituir el banco de reactivos o tareas evaluativas de un instrumento, esto implica la instauración de los comités académicos de elaboración de reactivos o de tareas a partir de la tabla de especificaciones. Posteriormente, con la ayuda de expertos, se efectúa el proceso de validación para eliminar en la medida de lo posible la varianza irrelevante al constructo y se analiza el contenido de las tareas evaluativas o los reactivos. Las acciones realizadas en este paso deben aportar evidencia que sustente que el instrumento mide lo que debe medir y que las inferencias que se realicen sobre los resultados son válidas.

3.1 **Desarrollar las tareas evaluativas o los reactivos de acuerdo con las especificaciones.** Este proceso debe apegarse de forma estricta a la tabla de especificaciones (estructura y especificaciones de reactivos y tareas evaluativas) y a los lineamientos técnicos. El elaborador debe comprender cómo están construidas las especificaciones y por qué es importante apegarse a ellas para desarrollar los reactivos y las tareas evaluativas (Haladyna, 2004). Es importante contar con un grupo de expertos amplio, plural y heterogéneo para el desarrollo de nuevos reactivos o de tareas evaluativas, a fin de garantizar su representatividad y prevenir algún tipo de sesgo en los productos.

3.1.1 En el caso de pruebas con reactivos de opción múltiple, se debe argumentar cada una de las opciones de respuesta (tanto la correcta como las incorrectas).

3.2 **Capacitar al comité elaborador de reactivos o tareas evaluativas**. Una vez instaurado el cuerpo colegiado que se encargará de la redacción de las tareas evaluativas o de los reactivos, debe ser capacitado, con el propósito de darle a conocer los insumos y los lineamientos técnicos necesarios para desarrollar adecuadamente su función.

La parte fundamental de la capacitación es la relacionada con los lineamientos técnicos a seguir, pues se explica cómo la construcción del reactivo impacta en los estadísticos que se obtendrán una vez administrado (Haladyna y Rodríguez, 2013 y Schmeiser y Welch, 2006). Se deben realizar sesiones teórico-prácticas (talleres) en las que se presenten ejemplos y se lleven a cabo ejercicios donde se apliquen los lineamientos técnicos a seguir.

3.3 **Atender los lineamientos técnicos para la construcción de reactivos**. Cada reactivo debe satisfacer un conjunto de reglas técnicas para su elaboración (calidad de imágenes, extensión de textos, etcétera), lo cual favorece la estandarización del instrumento y la obtención de estadísticos adecuados en la fase de piloteo. Estos lineamientos deben estar fundamentados en literatura especializada y, para el caso del tiempo de resolución, haber corroborado, de manera empírica, la factibilidad de su administración a la población objetivo.

3.3.1 En el caso de cuestionarios, se recomienda que la escala no incluya un punto medio, pues se corre el riesgo de que sea el intervalo que más se conteste debido a la deseabilidad social, además es una manera de forzar a las personas que contestan el instrumento a tomar una postura (Van Vaerenbergh y Thomas, 2012). Adicionalmente, todos los intervalos deben tener una etiqueta que los describa, con el fin de clarificar el significado de la escala.

3.3.2 En el caso de cuestionarios, todos los reactivos que pertenezcan a una misma escala se deben presentar en el mismo sentido. Sin embargo, es recomendable presentar de manera inversa algunas escalas a lo largo del instrumento para reducir determinados estilos o patrones de respuesta en los resultados y, por consiguiente, aminorar el sesgo (Buckley, 2009), siempre y cuando esto no implique problemas en la redacción de los reactivos, por ejemplo, presentar dobles negaciones.

3.3.3 En el caso de cuestionarios, en función del propósito de la evaluación, a fin de que haya mayor probabilidad de que se obtengan respuestas libres de deseabilidad social y que la evaluación sea menos amenazante para quien responde, se debe evitar incluir reactivos que de manera directa involucren la autoimagen de la persona evaluada (por ejemplo, parafraseando los reactivos en tercera persona) y, cuando sea posible, que el cuestionario sea contestado de manera anónima (Paulhus y Reid, 1991).

3.4 **Revisar técnicamente los reactivos o tareas evaluativas antes de ser validados.** Personal del organismo evaluador debe analizar los reactivos o las tareas para garantizar que éstos cumplan con los lineamientos técnicos que guiaron su elaboración. Si como resultado de la revisión técnica se identifica que es necesario algún ajuste, se debe recurrir al elaborador para realizarlo, también es posible presentarle propuestas y solicitar su apoyo cuando se requiera alguna precisión en el contenido del reactivo o la tarea evaluativa (Downing y Haladyna, 2006).

3.5 **Someter a revisión y corrección de estilo los reactivos o las tareas evaluativas.** Se debe verificar que el reactivo o la tarea no tenga errores ortográficos ni de redacción, con el fin de asegurar el uso adecuado del lenguaje y garantizar una comunicación eficaz. Es preciso definir criterios editoriales en un manual, que permitan dar unidad editorial a los instrumentos de evaluación; estos criterios deben ser conocidos por las personas que participan en la construcción de los cuestionarios y pruebas y su revisión tales como, elaboradores de reactivos y coordinadores de los instrumentos de evaluación, entre otros (INEE, junio 2016).

3.6 **Capacitar al comité de validación de reactivos o tareas.** Una vez instaurado el cuerpo colegiado que se encargará de la revisión de las tareas evaluativas o los reactivos, debe ser capacitado, con el propósito de darle a conocer los insumos y los lineamientos técnicos necesarios para desarrollar adecuadamente su función. Se debe revisar el propósito de la validación, el procedimiento a seguir y los dictámenes que podrán asignar, haciendo hincapié en la importancia del proceso y su repercusión en la calidad del instrumento. Adicionalmente, se debe ahondar en cada uno de los criterios que componen la lista de verificación utilizada.

3.7 **Validar los reactivos o las tareas evaluativas elaboradas.** Un grupo de expertos debe revisar cada uno de los reactivos o tareas elaborados, se debe asegurar que quienes validen un reactivo o tarea en específico sean expertos diferentes al elaborador, no conozcan al autor del reactivo y que accedan al material de trabajo sólo en el área y los tiempos asignados. El propósito de este ejercicio es garantizar que los reactivos o tareas se apeguen a lo definido en la tabla de especificaciones, que no tengan errores conceptuales ni de sesgo, es decir, este procedimiento tiene como finalidad sustentar que los reactivos o tareas se alinean al objeto de medida y son adecuados para obtener información que permita hacer inferencias acertadas, también es útil para minimizar la intervención de variables irrelevantes al constructo medido (AERA, APA y NCME, 2014). Se debe guiar la revisión con una lista de verificación en la que se destaquen los aspectos en los que el comité debe poner mayor atención, tendrá que enfocarse primordialmente en el contenido de las tareas y reactivos. Si como resultado de este proceso el comité decide hacer ajustes mínimos, o algunas precisiones para mejorar el reactivo o la tarea, se deben realizar en el momento, de manera que sea el mismo comité el que avale la última versión.

3.7.1 En el caso de pruebas, se debe garantizar que los reactivos no presenten más de una respuesta correcta y que ésta se encuentre asignada adecuadamente. Si es necesario reelaborar el reactivo o la tarea, debe descartarse y no pasar a la siguiente etapa. Este proceso debe ser sistemático y estandarizado (Schmeiser y Welch, 2006), de tal forma que los validadores dejen evidencia de su aprobación en la última versión de los reactivos o tareas.

3.7.2 En el caso de cuestionarios, se debe verificar que los reactivos de una misma escala se presenten en el mismo sentido. Si se optó por intercalar el sentido de las escalas, se debe revisar que los reactivos no presenten problemas en su claridad o interpretaciones (por ejemplo, que impliquen dobles negaciones).

3.7.3 En el caso de cuestionarios y escalas de autoevaluación, se debe incluir dentro de los criterios de validación la verificación de que los reactivos y las escalas recolectarán información libre de deseabilidad social y se presenten de tal forma que demanden a los participantes a tomar una postura clara sobre el aspecto que se está abordando. Si se decidió incluir una escala de deseabilidad social o viñetas de anclaje[9] para controlar el sesgo, se debe garantizar que cumplen con el propósito para el cual fueron hechas.

3.8 **Resguardar las tareas evaluativas o los reactivos en formato digital en un repositorio seguro.** Todas las tareas evaluativas o los reactivos elaborados, revisados y validados se deben resguardar en formato digital en un repositorio seguro que facilite su identificación y seguimiento en los subsecuentes procesos, que permita el registro puntual de quien los elaboró, los revisó y los validó, y que almacene los estadísticos obtenidos una vez que se haya administrado el instrumento (Haladyna y Rodríguez, 2013).

[9] Las viñetas de anclaje se componen de descripciones de personajes imaginarios con distintos niveles del constructo que se desea medir. Sirven para proporcionar un punto de referencia común para los evaluados con distintas preferencias en el uso de la escala, por ejemplo, aquéllos que tienden a sobrevalorarse. Las personas que contestan el instrumento identifican el nivel de rasgo de los personajes imaginarios con las mismas opciones de respuesta que emplearon para autoevaluarse. Con los resultados obtenidos, se ajusta el sesgo de la medición que se debe a las preferencias del uso de la escala de autoevaluación con la finalidad de obtener una estimación real del rasgo buscado. Para llevar a cabo este trabajo se parte de dos supuestos, el primero es que exista consistencia en las respuestas de las personas que contestan el instrumento, el segundo supuesto exige que haya equivalencia entre las viñetas, es decir, que sean entendidas de la misma forma por todas las personas (King, Murray, Salomon, y Tandon, 2004).

Evidencias para la verificación documental

- ✓ **Reactivos, tareas evaluativas o escalas elaboradas, revisadas y validadas**. Se resguardan los reactivos y sus datos de identificación en la plataforma destinada para ello, además de los reportes firmados por los comités. Esto debe servir para llevar un control de los procesos y para sustentar la correcta construcción del instrumento. Una vez pasada la validación y las revisiones descritas, los reactivos no tendrán errores técnicos, de contenido, de redacción o editoriales.
- ✓ **Bitácoras que documentan el trabajo de los comités académicos**. Las sesiones de trabajo serán respaldadas con las listas de asistencia y documentos en los que se describan las actividades desarrolladas en cada reunión, así como con los acuerdos firmados por los miembros de cada comité.
- ✓ **Currículum Vítae de los integrantes de los comités académicos de elaboración y validación de reactivos**. Este documento sustenta que las personas que participaron en este paso son expertas en el contenido del instrumento y tienen una trayectoria profesional que les permite desarrollar las actividades propias del comité.
- ✓ **Material de capacitación de los elaboradores y validadores**. Se debe presentar la carta descriptiva de la inducción de los comités académicos que participan en este paso, además de los ejercicios realizados para verificar la comprensión de los temas abordados.

Paso 4. Piloteo de las tareas evaluativas o de los reactivos

La intención del piloteo es probar empíricamente las tareas evaluativas o los reactivos, así como la logística de administración de los instrumentos. El piloteo debe realizarse utilizando una muestra de la población o subpoblaciones objetivo. Dicha muestra podrá ser probabilística o a conveniencia, en ambos casos se debe verificar que la selección de los elementos sirva a los propósitos del piloteo. En este paso se detectan posibles errores, defectos u omisiones procedimentales que pudieron pasar desapercibidos en la planeación general del instrumento (Paso 1), para que puedan ser ajustados antes de su administración (AERA, APA y NCME, 2014). Por ejemplo: tiempo de resolución de los reactivos y el instrumento en su conjunto, instrucciones, logística de la administración, entre otros. En ningún caso los resultados del piloteo serán usados para asignar puntuaciones a las personas que lo responden.

4.1 **Determinar la estrategia de piloteo de acuerdo con las características de la evaluación.** A partir del contexto de la evaluación, se debe establecer el procedimiento para poner a prueba los reactivos o las tareas evaluativas a fin de estimar su comportamiento estadístico.

4.2 **Seleccionar la muestra de personas que participará en el piloteo de acuerdo con los recursos disponibles.** Los individuos que participarán en el piloteo deben ser representativos de la población objetivo. Se deben proveer los detalles del procedimiento estadístico utilizado para establecer el tamaño de la muestra y la manera de seleccionarla (tipo de muestreo), así como para elegir a las personas que participarán. En caso de aplicar el muestreo probabilístico, también se deben documentar los niveles de confianza, el error de estimación y la tasa esperada de no respuesta utilizados para el cálculo del tamaño de muestra (INEGI, 2010). Si a partir de la muestra se van a estimar estadísticos, el tamaño de muestra tiene que ser suficiente para obtener estadísticos estables y con la precisión suficiente que permita tomar decisiones bien informadas sobre la inclusión, ajuste y eliminación de tareas evaluativas o reactivos que integrarán los instrumentos.

Es necesario incluir en la muestra una cantidad de individuos de cada una de las subpoblaciones que conforman la población objetivo que permita hacer estudios comparativos, a fin de detectar posibles sesgos en los reactivos o tareas.

4.3 **Incorporar a grupos con discapacidad al piloteo.** De haber adaptaciones del instrumento, deben participar en este proceso personas a quienes van dirigidas dichas adaptaciones en cualquiera de sus formatos: braille, letra grande, cinta de audio, computadora, presentación oral, entre otros (Thompson, Johnstone, Thurlow, y Altman, 2005; Schmeiser y Welch, 2006). Si por razones de logística esto no es viable, se pueden emplear técnicas cualitativas (grupos focales o entrevistas cognitivas, por ejemplo) para evaluar su pertinencia.

4.4 **Establecer el modelo de medición con el que se analizarán los datos del piloteo de acuerdo con el tamaño de la muestra.** Se debe especificar cuál será la metodología utilizada para el análisis de los datos. En todos los casos, se debe recopilar información sobre los parámetros que den cuenta de la calidad de los reactivos o tareas y, si el tamaño de la muestra lo permite, los posibles sesgos de medición. En principio, es mejor adoptar el mismo modelo de medición indicado en el diseño del instrumento (Paso 2); sin embargo, ciertos factores, como el tamaño reducido de la muestra para el piloteo, pueden justificar el uso de un modelo alternativo o complementario.

4.5 **Recurrir a métodos cualitativos cuando por razones justificadas no sea posible un estudio cuantitativo**. Si existen razones de peso por las cuales no sea posible pilotear los reactivos o las tareas a la cantidad de sujetos que exige el modelo de medición seleccionado para la obtención de sus parámetros, es posible recurrir a métodos de recolección de información cualitativa, tales como grupos focales, paneles de expertos, entrevistas cognitivas, entre otros. Estos métodos cualitativos se pueden llevar a cabo aun cuando sea posible realizar un estudio cuantitativo, pues pueden ser un recurso para identificar fuentes de sesgo y apoyar a las modificaciones a los reactivos o tareas evaluativas. Se debe sistematizar la estrategia empleada, así como documentarla.

Evidencias para la verificación documental

- ✓ **Estrategia de piloteo.** Este documento debe detallar el número de reactivos y formas que se probarán, el procedimiento para seleccionar a las personas que participarán en el estudio y sus características, el tamaño de la muestra, la justificación de la estrategia empleada, así como la inclusión de grupos especiales y el marco de la medición utilizado. Si la estrategia incluye métodos cualitativos tales como grupos focales, paneles de expertos, entrevistas cognitivas, entre otros, se debe documentar el procedimiento efectuado.
- ✓ **Resultados del piloteo**. Presentar los estadísticos obtenidos de los reactivos o las tareas y el dictamen de su funcionamiento, indicando los que alcanzaron los parámetros adecuados establecidos previamente y que dan cuenta de su buen funcionamiento, así como los que fueron descartados por no alcanzarlos. Cuando el piloteo ponga a prueba estrategias logísticas de administración del instrumento, se debe documentar su resultado.
- ✓ **Justificación del empleo de una estrategia cualitativa como forma de piloteo**. Fundamentar la estrategia cualitativa que de forma alterna fue empleada como piloteo de los reactivos y tareas evaluativas, documentando rigurosamente el procedimiento seguido, la población con la cual se realizó, así como los resultados obtenidos que guíen la toma de decisiones.

Paso 5. Ensamble del instrumento

Con la información de los pasos anteriores se seleccionan los mejores reactivos o tareas para su inclusión en el instrumento, tomando en cuenta la tabla de especificaciones y los estadísticos que obtuvieron los reactivos o las tareas en el piloteo. En este paso se debe considerar el número de formas necesarias, la modalidad de la aplicación (papel y lápiz o computadora) y el orden en que se presentarán los reactivos. Además, se debe planear la revisión de las formas, así como garantizar su calidad, legibilidad e integridad.

5.1 **Determinar la regla de diseño.** Asentar en un documento los mecanismos que se seguirán para garantizar la equivalencia de las formas del instrumento en términos del modelo de medición seleccionado, así como las características del ensamble para asegurar la comparabilidad de los resultados.

5.2 **Ensamblar el instrumento respetando las especificaciones de contenido.** Cualquier forma del instrumento debe ser ensamblada con base en lo establecido en la tabla de especificaciones del instrumento (Paso 2).

5.2.1 En el caso de pruebas es necesario que todas las formas de un mismo instrumento compartan las mismas características técnicas y sean ensamblados con reactivos que tengan estadísticos similares. En las distintas formas del instrumento, las medias y desviaciones estándar de la dificultad, así como los índices de confiabilidad o información, deben ser equivalentes. Asimismo, se recomienda incluir un porcentaje de reactivos comunes con buenas características métricas en las diferentes formas de un mismo instrumento, éstos para fines de estabilidad. La cantidad no debe ser menor a 30% ni mayor a 50% del total de reactivos efectivos para calificar; los reactivos seleccionados como ancla[10] deben estar ubicados en la misma posición relativa dentro de cada forma, y quedar distribuidos a lo largo de todo el instrumento y en toda la escala de dificultad. Estos reactivos deben tener, necesariamente, características cuantitativas ideales, según el modelo de medición seleccionado, además de cumplir cabalmente con todos los lineamientos técnicos establecidos en la fase de construcción del banco de reactivos.

5.2.2 Se debe elaborar una tabla técnica del ensamble de las formas del instrumento que contenga la información de la conformación de las diferentes formas del instrumento, detallando el número de identificación de cada reactivo, su clasificación temática de acuerdo con la tabla de especificaciones, en cuáles formas aparece y sus estadísticos.

[10] Se le llama reactivos ancla al conjunto de reactivos comunes entre dos o más formas de un instrumento, utilizado con propósitos de equiparación.

5.3 **Establecer criterios editoriales que den homogeneidad a la presentación del instrumento.** Todos los reactivos o las tareas evaluativas que integran las formas de un instrumento de evaluación deben presentarse de manera homogénea, para ello se deben seguir criterios editoriales y de diseño bien definidos y verificar que al ensamblar las formas no existan errores ortotipográficos ni de diseño (Haladyna y Rodríguez, 2013).

5.4 **Realizar pruebas de impresión o visualización de los instrumentos de evaluación.** Antes de reproducir el cuadernillo o de implementar el instrumento en computadora, se debe revisar su legibilidad, que no tenga errores ortográficos ni de diseño (tamaño y nitidez) y que la posición de los reactivos sea correcta; en el caso de los instrumentos administrados en línea, se debe tener la seguridad de que todas las funciones del programa estén operando correctamente, que se registren las respuestas de las personas que contestan el instrumento y que no haya otros programas habilitados en el equipo de cómputo que pudieran interferir en su administración.

5.5 **Establecer medidas para el resguardo de los instrumentos ensamblados.** Se debe garantizar que los instrumentos de evaluación ensamblados se resguarden de manera digital en algún repositorio que garantice su confidencialidad y seguridad. Si el instrumento es impreso, se deben establecer los mecanismos de resguardo y seguridad del lugar donde se almacenan.

Evidencias para la verificación documental

- ✓ **Regla de diseño**. Documento que describa los criterios cuantitativos y cualitativos para la integración de las formas para la administración del instrumento. En este documento se define la estrategia para controlar las variaciones de los parámetros de los reactivos que componen las diferentes formas y para que cualitativamente sean equiparables.
- ✓ **Tabla técnica de ensamble de las formas**. Debe presentar las características cualitativas y cuantitativas de los reactivos que integrarán cada forma y apegarse a la tabla de especificaciones del instrumento.
- ✓ **Protocolo de revisión editorial de las formas**. Documento en el que se especifiquen los criterios editoriales y de formación que garanticen la presentación estandarizada de los cuadernillos del instrumento.
- ✓ **Evidencia de las revisiones.** Documentos firmados en los que se evidencie que se llevó a cabo la revisión editorial de las formas.
- ✓ **Evidencia de las pruebas de impresión o visualización**.
- ✓ **Protocolos de resguardo de la información**.

Fase 3. Administración y resguardo del instrumento de evaluación

La administración del instrumento es uno de los pasos importantes en una evaluación, el garantizar condiciones estandarizadas para que las personas contesten el instrumento, contribuye a la validez de la interpretación de los resultados y la equidad de la evaluación (ACT, 2016).

Paso 6. Administración del instrumento

Este paso incluye los aspectos que se deben atender durante la administración del instrumento: instrucciones únicas, límites de tiempo, condiciones ambientales, supervisión, seguridad del instrumento y de las personas que lo contestan, entre otros. Gran parte de la validez de las inferencias que se van a realizar depende de que la administración se realice bajo condiciones estandarizadas (AERA, APA y NCME, 2014).

6.1 **Verificar que la administración del instrumento se haga bajo los estándares establecidos.** Se deben respetar los estándares de administración del instrumento tal y como se concibieron en la Planeación general y evitar ajustes al formato de presentación previsto que puedan alterar la medición del constructo (ACT, 2016).

6.2 **Contar con un plan de organización y logística.** Todo proceso de administración debe considerar una proyección completa del evento, desde la distribución de las formas del instrumento a lo largo del territorio de administración, la transportación de materiales, la preparación de las sedes (aforo), la distribución de los administradores del instrumento y los supervisores, la solución de imprevistos, el acopio y destino final de los materiales, entre otros.

6.3 **Definir la estrategia de difusión para la aplicación del instrumento.** Para garantizar la eficiente aplicación y asistencia puntual de las personas que contestarán el instrumento es necesario difundir con oportunidad los pormenores de la aplicación: fecha, hora y lugar; los requisitos para el ingreso a la sesión, la duración, entre otros.

En el caso de pruebas, particularmente de alto impacto, la guía del sustentante debe:

- Ser veraz y estar apegada a la documentación que sustenta la elaboración del instrumento.
- Difundirse con tiempo suficiente antes de la administración de la prueba (AERA, APA y NCME, 2014).
- Retomar los datos más relevantes de la evaluación: propósito, población objetivo y alcances y limitaciones, mismos que deben estar consignados en la ficha técnica del instrumento.
- Describir el objeto de medida.
- Contar con material de consulta.
- Incluir ejemplos de los reactivos o las tareas evaluativas.
- Presentar los procedimientos, las reglas y las sanciones que norman la conducta de los sustentantes antes, durante y después de la administración del instrumento. Informar sobre los materiales que se podrán ingresar durante la aplicación, las sanciones en caso de suplantación, copia o cualquier otra conducta no deseada (JCSEE, 2003).
- Proporcionar orientación sobre el acceso a condiciones adaptadas o modificaciones para aquellos sustentantes con discapacidad.
- Explicar el modelo de calificación que se utiliza (incluyendo si se lleva a cabo algún tipo de escalamiento), la forma en que se presentan y difunden los resultados, los niveles de desempeño y, si corresponde, los dictámenes que se entregan.
- Estar libre de errores ortográficos, de redacción o de diseño, de tal forma que la información sea clara y su presentación favorezca su comprensión.
- Describir los materiales de apoyo, si corresponde.

6.4 **Establecer mecanismos que permitan conocer de manera anticipada los apoyos que se requerirán para atender a personas con alguna discapacidad**. Para cumplir con este objetivo es necesario contar con la información que permita planear la administración de manera adecuada, por ejemplo: cuando una persona sea notificada de que será evaluada o se registre para una evaluación, se le deberá preguntar si requiere alguna adaptación o modificación, a fin de contar con el tipo de material necesario o con alguna persona capacitada en ese apoyo en específico.

6.5 **Administrar el instrumento en espacios con condiciones adecuadas.** La administración del instrumento de evaluación debe efectuarse en espacios que reúnan las condiciones de comodidad, higiene, iluminación y ventilación, así como aquellas específicas que se deban atender dada la naturaleza de la evaluación y donde exista el mínimo posible de distracciones. Se debe cuidar que las instalaciones sean accesibles y seguras para el desplazamiento de los asistentes, que cuente con rampas y baños de fácil acceso para las personas con discapacidad, así como con vías para el transporte adecuados. También se deben prever situaciones de emergencia y riesgo (DOF, 30/04/2014).

6.6 **Elaborar un manual de administración del instrumento (protocolo).** Para garantizar la estandarización, la instancia responsable del desarrollo del instrumento debe distribuir entre los administradores un documento escrito que contenga las instrucciones generales y procedimientos que se deben realizar en la administración del instrumento; además, este documento debe especificar las cualidades y responsabilidades de todos los involucrados, así como incluir los lineamientos para la atención a las personas con necesidades especiales que contestarán el instrumento (DOF, 30/04/2014).

6.7 **Seleccionar a los responsables de la administración del instrumento.** El reclutamiento de los administradores del instrumento se debe desarrollar en función de un perfil previamente establecido. Se debe optar por personas que tengan las cualidades necesarias para realizar esta función, tales como el trato cordial hacia las personas que contestarán el instrumento, así como un estricto apego al protocolo establecido para la aplicación y resguardo de la seguridad de los materiales.

6.8 **Capacitar a los responsables de la administración del instrumento.** Es necesario instruir de manera formal al personal responsable de la administración sobre las características del instrumento, los lineamientos a seguir en la administración, así como el código de conducta que deberá respetar. Es conveniente reunir a los administradores en una o más sesiones previas a la administración del instrumento para explicitar las responsabilidades de cada uno y aclarar dudas.

La capacitación debe incluir una simulación del procedimiento para familiarizar a los administradores con las características de la administración (por ejemplo, ciertos formatos de respuesta o el uso de computadoras). Durante la práctica, es necesario monitorear y retroalimentar a los participantes y, si es necesario, repetir la práctica hasta que la persona que administrará los instrumentos cuente con las habilidades necesarias para llevar a cabo las tareas que se le confieren.

6.9 **Capacitar a los lectores, intérpretes y escribas, si corresponde**. Los interpretes o lectores deben expresarse con fluidez, tener experiencia y conocimientos básicos del instrumento y del servicio que prestarán. Para lograrlo, previamente el organismo evaluador se encargará de que estas personas reciban la capacitación correspondiente y sean evaluadas.

6.10 **Informar a las personas que responden el instrumento si hay adaptaciones del mismo.** Si existen variantes del instrumento para atender algunas necesidades específicas, cada evaluado debe estar informado al respecto, de modo que pueda optar por que se le administre la variante que atienda mejor a sus características (DOF, 30/05/2011).

6.11 **Establecer un protocolo de seguridad de la información.** Es fundamental contar con un documento en el que se describan de manera detallada los lineamientos para garantizar la integridad y confidencialidad de los materiales, así como de los resultados y la información de las personas que contestarán el instrumento (Schmeiser y Welch, 2006). Se deben desarrollar los lineamientos y las políticas de uso de los resultados de la evaluación en caso de fraude. Este documento debe describir el rol y las responsabilidades que cada uno de los involucrados en el diseño y desarrollo del instrumento tendrá antes, durante y después de la administración, con la finalidad de garantizar la seguridad del mismo.[11]

6.12 **Establecer la llegada anticipada a la sede de materiales y administradores del instrumento.** Los materiales y personal responsable deben estar con antelación en el lugar en el que se realizará la administración del instrumento para garantizar que se cumplan todos los requerimientos físicos y materiales indispensables para efectuar adecuadamente el proceso de evaluación.

6.13 **Garantizar la estandarización de la administración del instrumento.** Con el propósito de controlar la precisión de la evaluación, es deseable, de una administración a otra o de un grupo a otro, mantener constantes las condiciones bajo las cuales se responde el instrumento (lectura de instrucciones, límite de tiempo) a fin de asegurar que los resultados se deben al instrumento y no a variables ajenas. Cuando exista más de una forma del instrumento, éstas deben ser administradas bajo las mismas condiciones (AERA, APA y NCME, 2014).

6.13.1 En el caso de pruebas, el organismo evaluador debe procurar que todos los evaluados tengan la misma oportunidad de mostrar su rendimiento y que nadie tenga alguna ventaja.

6.14 **Documentar situaciones que puedan afectar la administración del instrumento.** Si se presenta alguna eventualidad que afecte los resultados de la evaluación, se deben registrar todos los detalles en una bitácora; los administradores deben contar con los conocimientos que les permitan afrontar imprevistos o saber a quién acudir para resolverlos.

6.14.1 En el caso de pruebas, identificar vías de fraude y tomar medidas para sancionarlo. Durante la administración del instrumento es necesario estar consciente de las posibles formas para cometer fraude (copia de respuestas, acceso ilícito a materiales, suplantación, extracción de materiales, entre otras). Se puede utilizar una clasificación para las amenazas de copia y otra clasificación para las amenazas de robo, así como de las medidas necesarias para evitarlas o sancionarlas. Asimismo, y en la medida de lo posible, se deben controlar los eventos que puedan distraer a las personas que contestan el instrumento y detener su ejecución, pues se corre el riesgo de que la evaluación no sea equitativa o que se afecte directamente la validez de la misma al incrementar la varianza irrelevante al constructo medido. Por ejemplo: cortes de energía, solicitar que las personas se reubiquen, variaciones en el ancho de banda, equipos inadecuados, entre otros[12].

---

[11] Las directrices del ITC recomiendan establecer una serie de figuras que se encarguen de la seguridad del instrumento como un comité de seguridad, director de seguridad, vigilantes, entre otros; también, recomiendan que el plan de seguridad especifique los derechos y responsabilidad de los evaluados. Véase International Test Commission (2014). *The Security of Tests, Examinations, and Other Assessments.*Final Version. V. 1.0. Recuperado de:
https://www.intestcom.org/files/guideline_test_security.pdf

[12] Downing, S. y Haladyna, T. (2006). *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

6.15 **Determinar las situaciones en las que se deben invalidar las puntuaciones**. Se debe tener claridad en los casos que sea conveniente anular las respuestas de alguna persona, así como el procedimiento para hacerlo.

Evidencias para la verificación documental

- ✓ **Protocolo de seguridad para la administración del instrumento**. Este documento debe describir las estrategias que se implementarán para resguardar la confidencialidad de las formas, así como considerar el traslado de los materiales y los controles para su manejo en la sede de administración. Si se emplea la computadora, se deben contemplar las acciones que se implementarán para evitar que, haciendo uso de algún dispositivo, se extraigan los reactivos que conforman el instrumento. En cualquier caso, el protocolo debe precisar las acciones para enfrentar alguna eventualidad que atente contra la seguridad del instrumento.
- ✓ **Material de capacitación de los administradores**. Se debe presentar la carta descriptiva de la inducción de los responsables de la administración con el desglose de los contenidos y el tiempo dedicado a la revisión de cada uno, además de incluir las actividades realizadas y la evaluación del curso.
- ✓ **Bitácoras de la administración del instrumento**. En caso de que durante la administración ocurriera algún imponderable que pudiera afectar los resultados de la evaluación, como la falta de apego al protocolo de administración, errores de impresión o visualización del instrumento o extracción del instrumento, se deberán registrar dichos sucesos en bitácoras para analizar si afectan el resultado de las personas que lo contestan.
- ✓ **Manual de administración del instrumento**. Debe contener la información que es indispensable que el personal de apoyo conozca, tales como el lugar, tiempo, modo de administración de los instrumentos.
- ✓ **Actas sobre las anomalías en el proceso de administración del instrumento**, para el caso en que las haya.

Evidencias adicionales para el caso de pruebas

- ✓ Guía para el sustentante
- ✓ Materiales de apoyo: lecturas, formularios, etcétera

Paso 7. Resguardo de materiales al finalizar la administración del instrumento

Es indispensable desarrollar criterios y procedimientos para que, una vez que se realice la administración del instrumento, se aseguren los materiales utilizados y se salvaguarde su confidencialidad.

7.1 **Realizar el protocolo de manejo del material empleado en la administración del instrumento**. Se debe establecer de manera formal el procedimiento y las políticas para la transportación de los materiales de la administración: cuadernillos, hojas de respuestas, bitácora de incidencias, entre otras; también se deben especificar las medidas de seguridad para el resguardo de los cuadernillos y su posterior destrucción, a fin de garantizar su confidencialidad. Si la administración es en línea, asegurarse de determinar los mecanismos de seguridad para resguardar la información de la persona que responde al instrumento y las condiciones de la administración en forma electrónica, así como garantizar que las respuestas se registren adecuadamente y que la información llegue completa y segura para su procesamiento.

7.1.1 En el caso de pruebas, para referencias futuras, se sugiere guardar de manera segura todo el material impreso y electrónico utilizado en la administración del instrumento. Esto es particularmente relevante para evaluaciones de alto impacto, pues permiten responder a posibles controversias. Los materiales y productos del proceso de evaluación deben quedar bajo resguardo por un periodo mínimo de un año después de haber entregado los resultados.

7.2 **Tomar medidas para el control de los materiales**. Si durante la devolución del material de administración se detecta la falta de alguno de éstos, es necesario notificar al responsable del instrumento para que tome medidas al respecto. Todo esto debe quedar documentado en la bitácora. Una vez transcurrido el periodo de resguardo, los materiales utilizados en la administración deben ser destruidos mediante un procedimiento formal que quedará documentado en el acta legal correspondiente.

Evidencias para la verificación documental

- ✓ **Protocolo para el resguardo de la información**. Documento que describe las acciones que se realizan para el almacenamiento del material de administración del instrumento y las políticas de resguardo y de acceso a esta información. Se debe establecer el periodo de resguardo y el procedimiento para su destrucción.
- ✓ **Inventario de resguardo de materiales**. Control de la cantidad de material que se tiene bajo resguardo, esta evidencia debe facilitar su localización.
- ✓ **Acta de destrucción de materiales (si corresponde)**. Debe contener la cantidad y descripción de los materiales triturados o eliminados, el procedimiento empleado y la relación de testigos del procedimiento.

Fase 4. Análisis de resultados del instrumento de evaluación

En esta fase se mencionan los elementos que deben considerarse para el análisis estadístico del instrumento y de los reactivos o las tareas evaluativas que lo conforman, así como el modelo de puntuación utilizado. Conocer las propiedades métricas de los reactivos permite identificar los que son aptos para ser considerados en la puntuación de los sujetos, así como aquellos que no aportan información relevante o son defectuosos, por lo que deben ser excluidos (Cook y Beckman, 2006; Downing, 2004; Stemler y Tsai, 2008).

Paso 8. Evaluación de la métrica del instrumento

En este paso se llevan a cabo los análisis estadísticos considerando las respuestas a los reactivos o las tareas evaluativas, con el objetivo de aportar evidencia para la validez de las inferencias que se desean obtener a partir de los datos recopilados. Esta información es de vital importancia, debido a que el propósito de este paso es fundamentar empíricamente las características del instrumento de evaluación.

8.1 **Establecer un protocolo para la lectura de la Hoja de respuestas (administración lápiz-papel) y de corrección de bases de datos.** Antes de iniciar el análisis estadístico, es conveniente contar con los procedimientos documentados para la digitalización y sistematización de las respuestas de las personas, además de los mecanismos que se emplean para depurar las bases de datos con el propósito de minimizar los errores de codificación.

8.2 **Definir un protocolo de análisis de la métrica del instrumento.** Se debe formalizar el procedimiento que se realizará para valorar la pertinencia estadística de los reactivos, tareas evaluativas y del instrumento integrado (opción múltiple o respuesta construida). Esta información debe ser congruente con el modelo de medición establecido en la planeación general del instrumento al inicio de su construcción. Además, se deben incorporar los valores deseados de cada parámetro e indicador y su justificación.

8.3 **Realizar los análisis de los instrumentos con base en el protocolo establecido**:

a) estimación de estadísticos de reactivos o tareas;

b) confiabilidad/función de información[13]/índice de generalizabilidad;

c) dimensionalidad;

d) funcionamiento diferencial (reactivos e instrumento).

8.3.1 Para el caso de cuestionarios, verificar que la medición no esté influenciada por la deseabilidad social. En este sentido, una de las técnicas o procedimientos que se emplean con más frecuencia es utilizar una escala explícita de deseabilidad social en conjunto con la escala de interés y verificar si existe una asociación significativa y moderada entre ellas; de ser así, se asume la presencia de deseabilidad social (Havercamp y Reiss, 2003)[14].

8.3.2 En cuanto a los cuestionarios, elegir y justificar el método para la corrección de los puntajes, en caso de detectar la presencia de algún estilo de respuesta, con el propósito de reducir el valor de la invarianza en la medida.

---

[13] Concepto en la Teoría de Respuesta al Ítem (TRI) que indica la precisión con la que el reactivo estima el grado de constructo latente que posee el objeto evaluado a lo largo de la escala. En el caso de pruebas, un reactivo o tarea evaluativa proporcionará mayor información sobre los sustentantes cuyo nivel de habilidad esté cerca de su dificultad, que de los sustentantes con una habilidad alejada de este punto.

[14] Estadísticamente, existen otros métodos para detectar la deseabilidad social y controlarla, como el modelo de observaciones espurias, el modelo de supresión o, el modelo moderador (Ganster, Hennesey, y Luthans, 1983), análisis de factores confirmatorio, análisis confirmatorio de clases latentes, método del punto medio (*midpoint responding*), respuestas extremas (*extreme responding*), entre otros (Van Vaerenbergh y Thomas, 2012).

8.4 **Obtener el índice de respuestas omitidas**. Independientemente del marco de la métrica, es recomendable calcular el índice de no respuesta para el reactivo o la tarea evaluativa, es decir, la proporción de las personas que no lo contestaron. Un análisis exhaustivo de esta información ayuda a decidir qué tratamiento dar a los reactivos sin respuesta en la puntuación del instrumento y reconsiderar el tiempo de administración.

8.5 **Verificar el ajuste de los datos, en el caso de la TRI.** Es importante incluir la valoración del ajuste del comportamiento de los reactivos al modelo de medición establecido. Se recomienda llevar a cabo contrastes para evaluar la bondad de ajuste de cada reactivo[15]. Es decir, es recomendable no sólo considerar la estimación puntual de cada índice, sino también su precisión, particularmente, en caso de que el tamaño de la muestra sea pequeño.

8.6 **Estimar el error de medida**. Cualquier instrumento es vulnerable a todas las fuentes de error que a veces causan fluctuaciones en los resultados de una fuente de información a otra o de un área de medición a otra. Debido a que hay diversas fuentes de error, hay diferentes procedimientos para estimar la confiabilidad, por ejemplo: consistencia interna, estabilidad temporal y concordancia entre los evaluadores. Por lo que se debe elegir el más apropiado al tipo de instrumento que se construyó. La forma para obtener las evidencias de confiabilidad: índice de generalizabilidad, la función de información, error estándar condicional, índice de consistencia, entre otros, debe ser apropiada para los usos previstos de los resultados, la población involucrada y los modelos de medición utilizados para obtener las puntuaciones. Con el fin de conocer qué tan precisa es la medición y cuáles son sus principales fuentes de varianza se emplean la confiabilidad en el caso de la Teoría Clásica de los Test, la función de información en el caso de la Teoría de la Respuesta al Ítem o el índice de generalizabilidad para el caso de la Teoría de la Generalizabilidad[16].

Un mayor grado de precisión es requerido para evaluaciones de alto impacto. A la inversa, un grado más bajo puede ser aceptable cuando una decisión con base en la puntuación de un instrumento es reversible o depende de su corroboración de otras fuentes de información.

8.7 **Realizar análisis de funcionamiento diferencial de los reactivos y del instrumento.** Los análisis DIF y DFT deben hacerse considerando grupos de interés donde se sospeche que los reactivos pueden funcionar de manera diferente, si esto llegara a ocurrir se recomienda que el reactivo no sea considerado para emitir los resultados; algunas variables que se pueden considerar para estos análisis son: género, creencia religiosa, lengua materna, estrato socioeconómico y grupo étnico. El aspecto más relevante de este tipo de análisis es fundamentar que la prueba, en su conjunto, no afecta diferencialmente a los sujetos debido a su grupo de pertenencia (Muraki, 1999; Fidalgo y Madeira, 2008).

8.7.1 Para los cuestionarios, se debe identificar si existen diferencias transculturales en el uso de las escalas Likert o en reactivos categóricos individuales extraídos de esas escalas. Si se encuentra que alguna subpoblación muestra algún estilo de respuesta específico para responder, dadas sus características culturales, se debe reportar y considerar en el momento de comunicar las inferencias de los resultados. Los posibles estilos o patrones de respuesta son: sesgo de positividad, es una tendencia a estar de acuerdo con los reactivos independientemente de la actitud real; estilo discrepante, es una tendencia a no estar de acuerdo con los elementos, independientemente de su contenido; estilo de respuesta extrema, se refiere a elegir los puntos más lejanos de la escala y, por último, estilo de respuesta no contenciosa, se emplea para describir la selección aleatoria de las respuestas de los reactivos (Buckley, 2009).

8.8 **Dar a conocer la precisión de la medición con oportunidad**. Los resultados de los análisis para determinar el error de medida o la consistencia de la medición se deben reportar para que sean considerados al momento de la interpretación de los resultados, y ésta sea adecuada (JCSEE, 2011 y AERA, APA y NCME, 2014).

---

[15] Al interpretar los índices de los reactivos se debe tomar en cuenta que son estimaciones de parámetros poblacionales propensas a errores por fluctuaciones muestrales.

[16] La confiabilidad informa cuánto influyen en la estimación los errores aleatorios, no sistemáticos; una confiabilidad baja refleja mucho error de medición y, por lo tanto, las inferencias que se hacen a partir de la estimación son menos precisas y, por ende, tendrán menos validez. Del mismo modo, la función de información indica la precisión de la estimación a lo largo de la escala de habilidad, es decir, para niveles de habilidad distintos, la precisión con la que se mida también será diferente.

8.9 **Dictaminar la pertinencia de los reactivos y las tareas evaluativas con base en sus parámetros.** Con base en el modelo psicométrico seleccionado y los estándares establecidos en el protocolo, se deben revisar los resultados del análisis cuantitativo para determinar cuáles reactivos y tareas evaluativas son métricamente adecuados; este ejercicio consiste en comparar los índices obtenidos con los valores deseados, lo que debe dar como resultado una etiqueta que resuma su funcionamiento.

8.10 **Integrar la información cuantitativa al banco de reactivos y tareas**. Los parámetros o indicadores obtenidos en este paso deben ser incorporados, de manera periódica, en el repositorio destinado para el resguardo de reactivos, de modo que esta información pueda ser consultada en futuros ensambles y en el mantenimiento del instrumento.

8.11 **Determinar la inclusión en los análisis estadísticos de las respuestas de personas que contestaron un instrumento adaptado o de aquéllas a las que se hizo alguna modificación en las condiciones de administración**. Según el impacto de la evaluación y sus características, se debe hacer un ejercicio de reflexión para determinar, a la luz del principio de equidad (UNICEF, 2012), si las respuestas de estas personas son incluidas en los análisis con el resto de los resultados. El producto de esta reflexión debe hacerse explícito en el protocolo correspondiente. Con el propósito de facilitar esta decisión se propone apoyarse en la siguiente clasificación (Lewis, Patz, Sheinker y Barton, 2002):

Categoría 1. Se definen como adaptaciones y modificaciones que no se espera que influyan en el desempeño de la persona que contesta el instrumento de una manera que altere la interpretación de los resultados de la evaluación. Se caracterizan por ser ajustes que no alteran la medición ni sus condiciones de manera directa. Por ejemplo: magnificar el tamaño de letra del instrumento o el sonido en los estímulos auditivos; destinar a las personas un espacio exclusivo para la presentación del instrumento o tomar descansos adicionales sin que esto signifique aumentar el tiempo de administración.

Categoría 2. Se definen como adaptaciones y modificaciones que pueden tener un efecto en el desempeño de las personas y que deben ser consideradas al interpretar su puntuación. Es posible que mediante estudios cuantitativos se muestre que este tipo de ajustes no se comporta estadísticamente diferente, en comparación con el resto. En caso contrario, los resultados del análisis deberán interpretar con esa salvedad. Un ejemplo de esta categoría es aumentar el tiempo de administración, leer en voz alta, resaltar con negritas las instrucciones, entre otras.

Categoría 3. Se definen como adaptaciones y modificaciones que de alguna manera modifican el objeto de medida y pueden alterar la interpretación de los puntajes obtenidos. Esto ocurre cuando el ajuste está fuertemente relacionado con el constructo o la variable latente que se está midiendo (por ejemplo: tener una prueba de comprensión de lectura leída en voz alta). A falta de una investigación que muestre lo contrario, los resultados de los instrumentos y las consecuencias o decisiones asociadas con ellos, deben interpretarse no sólo a la luz de los ajustes realizados, sino también a la luz de cómo éstos pueden alterar lo que se mide. Ejemplos de esto son la traducción de un instrumento a otro idioma o a Braille, el uso exclusivo de un diccionario o el empleo de un lector o un escriba.

8.12 **Realizar los análisis estadísticos a las formas adaptadas.** Se debe determinar el modelo de medición a partir del cual se realizarán los análisis de este tipo de formas. Este criterio incluye la equiparación de puntajes para que los resultados de todas las formas sean comparables y la identificación de los aspectos o componentes del instrumento que pudieran ser equivocadamente evaluados para alguna población (Kolen y Brennan, 2014). Si las puntuaciones totales se presentan por separado para los grupos evaluados, se debe valorar la comparabilidad de las puntuaciones (adaptaciones contra población total). Si esta evidencia indica que existen diferencias entre los grupos, se debe examinar la validez de las interpretaciones de los resultados y proporcionar declaraciones preventivas.

8.13 **Documentar las características de las traducciones (si corresponde).** Si se hace una traducción del instrumento a otro idioma o lengua, se debe describir el proceso y evaluar los resultados y su comparabilidad con la versión original.

Evidencias para la verificación documental

- ✓ **Protocolo de lectura de las hojas de respuestas y corrección de las bases de datos**. Describe los mecanismos para garantizar una base de datos libre de errores de codificación.
- ✓ **Protocolo del análisis de la métrica utilizada**. Establece los métodos para estimar las propiedades cuantitativas de los reactivos, de las tareas y del instrumento como una unidad; debe ser congruente con el alcance de la evaluación y con el modelo estadístico seleccionado.
- ✓ **Resultados de los análisis realizados**:
  - a) estimación de parámetros
  - b) confiabilidad/función de información/índice de generalizabilidad;
  - c) dimensionalidad;
  - d) funcionamiento diferencial (reactivos e instrumento).
- ✓ **Reactivos y tareas dictaminados.** El banco donde se resguardan los reactivos y las tareas del instrumento debe almacenar los estadísticos, indicadores y las etiquetas que dan cuenta de su funcionamiento.
- ✓ **Si se realizaron adaptaciones al instrumento, reportar los resultados obtenidos.**

Evidencias adicionales en el caso de cuestionarios o instrumentos con escalas

- ✓ Protocolo de los análisis realizados para controlar la deseabilidad y su interpretación.
- ✓ Protocolo para la identificación de patrones o estilos de respuesta y su tratamiento para aminorar el efecto en los resultados de la evaluación.

Paso 9. Resultados de las puntuaciones obtenidas a partir del instrumento

Este paso se refiere a la asignación de un valor a las respuestas de las personas en cada reactivo o tarea evaluativa del instrumento, a fin de que al combinarlos se obtenga una o más puntuaciones, con base en el modelo establecido en la fase de planeación (AERA, APA y NCME, 2014).

9.1 **Elaborar el protocolo para obtener la puntuación de los instrumentos.** La manera para derivar las puntuaciones de las personas que responden el instrumento debe ser descrita con suficiente detalle en un protocolo para garantizar su exactitud y pertinencia técnica. Se debe especificar cómo se trabajará el modelo de puntuación, por ejemplo, si se trata de una escala que se puntúa de manera dicotómica o politómica. Cuando se incluyan procesos de jueceo se deben documentar con claridad los procedimientos y criterios que se utilizarán para asignar la puntuación a las personas que contestaron el instrumento.

9.1.1 En el caso de pruebas, el protocolo para obtener la puntuación debe considerar:

- Validar las claves de respuesta. El primer paso para calificar un instrumento es verificar las claves de respuesta correcta en los reactivos de opción múltiple. En el caso de instrumentos de respuesta construida, se debe revisar que el protocolo para la asignación de las puntuaciones favorezca a que la rúbrica de calificación se emplee de manera adecuada, además de verificar que no haya errores en el registro. Este criterio es obligatorio y debe quedar constancia de que se implementó.
- Obtener, en el caso de los instrumentos de respuesta construida en los que el desempeño es evaluado a través de una rúbrica, el porcentaje de acuerdos interjueces e intrajueces antes de dar la calificación definitiva a las personas que contestaron la prueba. Se deben tener previstas las acciones que se implementarán en caso de haber una discordancia mayor a la tolerada (considerando un criterio establecido previamente), por ejemplo: reforzar la capacitación de los jueces o prescindir de los servicios de otros cuando existan decisiones consistentemente discordantes.
- Establecer criterios para seleccionar a los evaluadores de los instrumentos que requieran la calificación de jueces. Para calificar las tareas evaluativas por medio de una rúbrica se deben considerar, al menos, dos evaluadores y es indispensable establecer criterios de selección: el perfil académico y profesional, el dominio de la tarea y, en caso que haya participado en el proyecto anteriormente, los porcentajes de acuerdo inter o intrajuez (Jonsson y Svingby, 2007; Rezaei y Lovorn, 2010; Stemler y Tsai, 2008; Stellmack, et. al, 2009).

- Se deben establecer los mecanismos de seguridad para que el personal que realice la calificación sólo tenga acceso a las respuestas de los sustentantes sin conocer ninguno de los datos de identificación; por ejemplo: nombre, Registro Federal de Contribuyentes (RFC), Clave Única de Registro de Población (CURP). Dichos mecanismos asegurarán el anonimato de las respuestas de los sustentantes.
- Capacitar y monitorear a los evaluadores del instrumento de respuesta construida. Se debe capacitar a los evaluadores para que se familiaricen con la rúbrica y para unificar sus decisiones (Council of Europe, 2011). También se proporcionará evidencia del grado de acuerdo entre las puntuaciones dadas por los jueces, a fin de que la calificación obtenida sea un resultado imparcial y objetivo de la medición.

9.1.2 En cuanto a los cuestionarios, el protocolo para obtener la puntuación debe considerar el tratamiento que se le dará, en caso de detectar algún estilo de respuesta (extrema, respuestas intermedias o sobrevaloración), con el fin de controlar la deseabilidad social y aumentar la validez de las inferencias que se realizan a partir de los resultados de la evaluación.

9.2 **Calcular las puntuaciones de las personas que responden el instrumento de acuerdo con el algoritmo determinado.** El cálculo de las puntuaciones debe hacerse de acuerdo con lo establecido en el diseño del instrumento. Los análisis estadísticos pueden motivar cambios en el esquema de puntuación utilizado; sin embargo, cualquier desvío del plan original debe documentarse y justificarse.

9.2.1 En el caso de pruebas, escoger el método de equiparación más adecuado, si las condiciones del instrumento lo permiten[17]. Si debido a la exigencia de las reglas para equiparar no se pudo llevar a cabo dicho proceso, entonces se deben escalar los puntajes de las diferentes formas para garantizar equidad en la evaluación[18].

9.2.2 En el caso de pruebas, si el reporte de calificaciones incluye la clasificación de los evaluados en niveles de desempeño, se debe documentar el procedimiento empleado para el establecimiento de puntos de corte, así como el método seleccionado (Angoff, Beuk, Bookmark, Nedelski, Hofstee, grupos de contraste, etcétera [INEE, 2016 y Downing y Haladyna, 2006]) y su justificación. Los descriptores de los niveles de desempeño y los puntajes asociados deben estar fundamentados y conocer el error estándar de medida asociado a las puntuaciones e incorporarlo en el reporte técnico del instrumento.

9.2.3 En el caso de instrumentos de autoevaluación, si se incluyeron viñetas de anclaje o alguna otra estrategia para el control de la deseabilidad social o el estilo de respuesta, se debe realizar el escalamiento de los resultados a través del método seleccionado (paramétrico o no paramétrico) (King y Wang, 2006).

9.3 **Verificar que, a partir de los instrumentos adaptados, se realizan inferencias similares al resto de las formas.** Las adaptaciones de un instrumento deben producir conclusiones equivalentes a las generadas con las formas no adaptadas, por lo que las inferencias que se pueden derivar de ellas se deben justificar con evidencias o estudios empíricos (ETS, 2015).

Evidencias para la verificación documental

- ✓ **Protocolo de puntuación.** Contiene los algoritmos y las reglas de puntuación establecidos en la ficha técnica del instrumento y la manera correcta de interpretar los resultados.
- ✓ **Evidencias de los estudios empíricos** que se realizaron y que sustenten que las inferencias de los instrumentos adaptados se corresponden con los instrumentos originales.

Evidencias adicionales en el caso de cuestionarios o instrumentos con escalas

- ✓ Protocolo para realizar el escalamiento de los resultados, incorporar evidencias adicionales en el caso de haber utilizado viñetas de anclaje o alguna otra estrategia para el control de la deseabilidad social o el estilo de respuesta.

[17] La equiparación 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. Ajusta, por dificultad, las distintas formas que fueron ensambladas bajo las mismas reglas estadísticas y de contenido (Holland y Strawderman, 2011 y Kolen y Brennan, 2014).

[18] El escalamiento se lleva a cabo a partir de las puntuaciones crudas (cantidad de aciertos) de los sustentantes, y se obtiene una métrica común para todos los instrumentos de evaluación (Shun-Wen, 2006 y Wilson, 2005).

Fase 5. Difusión, uso y resguardo de los resultados del instrumento de evaluación

El aspecto central de esta fase es la adecuada interpretación y difusión de los resultados de la evaluación, de acuerdo con sus objetivos, alcances y limitaciones. Los reportes de resultados pueden elaborarse para las diferentes audiencias a las que está dirigida la evaluación (ETS, 2015; ITC, 2014).

Paso 10. Reportes y uso de los resultados

Los medios que se emplean para dar a conocer los resultados de la evaluación deben cumplir con ciertos requisitos para garantizar la validez de las interpretaciones que se hagan de dichos resultados, así como asegurar su imparcialidad y que la información que se brinde a los usuarios sea oportuna y adecuada. La instancia que desarrolla el instrumento está obligada a realizar propuestas de comunicación e interpretación de los resultados derivados de la evaluación, alertando sobre los límites, alcances y limitaciones de la misma (AERA, APA y NCME, 2014).

10.1 **Establecer la escala de referencia para la difusión de los resultados.** De acuerdo con el propósito de la evaluación, el tipo de instrumento y la experiencia de los usuarios, a fin de hacer más sencilla y fluida la difusión de los resultados, se debe determinar la escala que se empleará para tal efecto: porcentaje, puntuaciones escaladas, entre otras (Downing y Haladyna, 2006). Si las puntuaciones del instrumento son una transformación de los puntajes crudos, es recomendable que esta escala sea de fácil interpretación y se describan sus características principales e implicaciones. Si existe el riesgo de que malinterpreten los resultados, se debe incluir una advertencia para evitar conclusiones erróneas. En la medida de lo posible, se debe presentar la información en un lenguaje sencillo, comprensible para todas las personas a las que está dirigida.

10.1.1 En el caso de pruebas, se debe elaborar el reporte de resultados para los evaluados, que contenga los siguientes elementos (JCSEE, 2011):

- estar redactado en un lenguaje que permita a los evaluados entenderlo;
- contener la información necesaria para comprenderlo;
- explicar la métrica empleada para la valoración del desempeño de los sustentantes;
- contener información que permita la interpretación adecuada de los resultados;
- difundir información confiable;
- no tener errores tipográficos ni de formato;
- no tener sesgos;
- entregarse con oportunidad.

10.2 **Elaborar reportes de resultados adicionales a los dirigidos a los usuarios de la evaluación.** Si la evaluación lo considera, se debe elaborar un reporte de resultados para alguna institución, dependencia o público en general con la finalidad de dar retroalimentación o rendir cuentas. Sin embargo, es indispensable que se advierta sobre los alcances de la evaluación, las interpretaciones que deben realizarse a partir de los resultados, así como el uso adecuado de los mismos, especialmente si hay condiciones específicas de la evaluación, como el que se haya realizado a la totalidad de la población, objeto de la evaluación, así como las consideraciones que se deben tener si se hizo un plan de muestreo (y las implicaciones de su incumplimiento) o, bien, si la participación de las personas que contestaron el instrumento fue casuística.

10.3 **Alinear cualquier reporte a la ficha técnica del instrumento.** Cualquier documento que proporcione resultados de la evaluación, debe considerar la información establecida en la ficha técnica que fundamentó el desarrollo de la evaluación como el propósito, alcance, uso de los resultados, entre otros.

10.4 **Recopilar evidencia que sustente de manera coherente las interpretaciones de los resultados**. Obtener y documentar la evidencia conceptual, teórica y empírica para sustentar que el instrumento cumple con el propósito para el cual fue elaborado, es decir, que obedece a los fines previstos y apoya las derivaciones de los resultados en la población objetivo. Esta evidencia debe ser coherente y completa para garantizar la idoneidad de las inferencias que se hacen y las acciones realizadas en consecuencia. Las inferencias que tengan mayor impacto deben sustentarse con más evidencia (Kane, 2013).

10.5 **Comunicar los límites y alcances de la evaluación.** Es responsabilidad de la instancia que desarrolla la evaluación, dar a conocer de manera explícita las limitaciones asociadas a la evaluación y relacionadas con el uso de las puntuaciones obtenidas. Esta información debe ser considerada para el fundamento de las decisiones que se tomen (AERA, APA y NCME, 2014).

10.5.1 En el caso de cuestionarios, declarar la presencia de estilos de respuesta y la estrategia empleada para su corrección; además del impacto en la interpretación de los resultados de la evaluación (He y Van de Vijver, 2016).

10.6 **Respetar siempre el propósito de la evaluación.** Una vez comunicados los resultados a partir de los criterios aquí presentados, por ningún motivo se deben emplear para tomar decisiones no previstas en la planeación general del instrumento. Realizar interpretaciones fuera de los límites de la evaluación es una amenaza para su validez (Downing y Haladyna, 2006).

10.7 **Orientar a los usuarios sobre los usos de las puntuaciones y los resultados de la evaluación.** Si es necesario, el organismo evaluador debe proporcionar asesoría a los beneficiarios de las puntuaciones o los resultados del instrumento que les ayude a recolectar e interpretar su propia evidencia de validez. Si por alguna razón las instancias, organismos o instituciones usuarias de la información desean emplear los resultados para fines distintos a los indicados expresamente por los referentes de la evaluación, ellas son las responsables de planear, realizar e interpretar los resultados de los estudios de validez locales (Cook y Beckman, 2006; ETS, 2007).

Evidencia para la verificación documental

- ✓ **Reportes de resultados (individuales o institucionales)**. Deben contar con las siguientes características:
  - ser consistentes con el propósito de la evaluación;
  - no tener errores tipográficos ni de formato;
  - contener la información necesaria para comprenderlos;
  - no tener sesgos;
  - difundir información confiable;
  - contar con información que permita interpretar adecuadamente los resultados;
  - estar redactados en un lenguaje que permita entenderlos;
  - presentarse con oportunidad;
  - explicar la métrica empleada.
- ✓ **Documento que comunique los límites, alcances de la evaluación, así como los usos previstos de la misma**.

Paso 11. Resguardo de la información

En este paso se desarrollan estrategias para la adecuada custodia del instrumento y de los resultados de su administración.

El banco de reactivos o tareas evaluativas es el repositorio donde son resguardados y clasificados. Se conserva información importante como las características métricas obtenidas en las diferentes aplicaciones, el tamaño de la muestra, el número de veces y fechas en que el reactivo o la tarea se integró a una forma, etcétera. En este banco, cada tarea evaluativa o reactivo cuenta con un identificador único y con un historial de diseño y uso, lo que permite valorar su calidad técnica (Ward y Murray-Ward, 1994).

El banco permite suministrar y almacenar los reactivos y las formas en un repositorio seguro, así como la participación de los diferentes usuarios (elaboradores de reactivos, validadores de reactivos, coordinadores de la prueba y revisores de estilo, entre otros).

11.1 **Contar con un banco de reactivos o de tareas evaluativas.** El banco de reactivos o de tareas evaluativas debe instalarse en un repositorio especializado y de fácil acceso que permita su manejo por los diferentes usuarios, así como su debida gestión. Sus elementos deben estar organizados y clasificados con base en el objeto de medida definido en la tabla de especificaciones. En el banco de reactivos, cada reactivo o tarea evaluativa debe contener, además de la información de los valores de los parámetros obtenidos, la clave de identificación, el tipo de respuesta (opción múltiple, abierta), ajustes o comentarios realizados por el revisor e historial del reactivo.

11.2 **Establecer protocolos de intercambio de información, en caso de que diferentes actores sean responsables de su procesamiento.** Con el propósito de hacer más efectiva la comunicación entre los diferentes responsables, es necesario documentar formalmente los lineamientos y las características de los entregables: programas informáticos, diccionarios de datos, bases de datos, reportes, entre otros; así como las funciones de cada uno de los participantes en el procesamiento de la información.

11.3 **Resguardar las respuestas de las personas que respondieron el instrumento y los resultados.** Es esencial que nadie, sin la autorización requerida, tenga acceso a la información original (las respuestas de las personas que contestaron el instrumento) ni a la información procesada (análisis estadístico y puntuaciones). Del mismo modo, al reportar los resultados se debe cuidar que la información se envíe y llegue a las personas correctas. En general, a pesar de que la confidencialidad y seguridad es responsabilidad compartida de todos los involucrados en el desarrollo y el uso del instrumento, se recomienda elaborar y aplicar protocolos para su protección.

11.4 **Delimitar el tiempo, la disponibilidad y el uso de los resultados.** Las organizaciones que conserven los resultados de los instrumentos de evaluación deben tener lineamientos explícitos en cuanto al tiempo que los conservarán, su disponibilidad pública y su uso a lo largo del tiempo. Lo anterior debe cumplir con lo establecido en la Ley General de Transparencia y Acceso a la información Pública (DOF, 04/05/2015) respecto a la clasificación de la información en sus modalidades de reservada o confidencial en el caso de datos personales; y conforme al plazo señalado como vigencia documental dentro del Catálogo de Disposición Documental establecido en el artículo 4, fracción X, así como las disposiciones secundarias correspondientes a la materia de archivos. Se deben proporcionar medidas de seguridad para proteger los resultados de la evaluación de las deformaciones por sentimientos personales y por sesgos en cualquier área de la evaluación. Esto también aplica para los informes o reportes de resultados (JCSEE, 2011).

Evidencias para la verificación documental

- ✓ **Banco de reactivos o de tareas evaluativas**. La herramienta debe almacenar los datos de identificación de los reactivos o de las tareas evaluativas, su historial, sus parámetros estadísticos actualizados y el registro de las personas que intervinieron en su construcción.
- ✓ **Protocolo para el resguardo de las bases de datos**. Documento que explique las políticas y el procedimiento para custodiar las bases de datos, así como las medidas de seguridad para proteger su integridad y el acceso a ellas.
- ✓ **Protocolo de intercambio de información**. En caso de que participen diferentes actores en el procesamiento de la información, este documento describe las responsabilidades de cada uno, los lineamientos para el proceso y las condiciones de los entregables, por ejemplo: diccionarios y bases de datos.
- ✓ **Bases de datos e informes de resultados.** Deben contener la información íntegra y no tener errores de contenido.

Fase 6. Mantenimiento del instrumento de evaluación

Si es necesario que el instrumento siga funcionando después de la primera administración, es indispensable considerar una fase de mantenimiento. A partir de los análisis realizados, se efectúa un estudio minucioso del instrumento para establecer un plan de trabajo que permita mejorar su calidad técnica, a fin de que se pueda utilizar en administraciones subsecuentes. Asimismo, se debe considerar la incorporación de reactivos o tareas evaluativas con estadísticos adecuados al banco.

Paso 12. Actualización del objeto de medida y del banco de reactivos o de tareas evaluativas

12.1 **Revisar el objeto de medida del instrumento de evaluación.** El contenido del instrumento y su vigencia deben ser revisados, a fin de verificar su pertinencia para cumplir el propósito de la evaluación, mediante un proceso de mantenimiento. Este procedimiento debe realizarse cuando se observen cambios en el objeto de medida (avance en la disciplina, transformación natural del objeto de medida, algún cambio en el alcance de la evaluación, entre otros). Es responsabilidad del organismo evaluador vigilar las condiciones bajo las cuales se desarrolla el instrumento, se modifica o se revisa; además de decidir cuándo es necesario retirarlo de operación (Muñiz, 2003).

12.2 **Realizar un diagnóstico inicial con toda la información disponible**. El organismo evaluador debe realizar análisis cualitativo y cuantitativo de la tabla de especificaciones, de los reactivos y las tareas evaluativas a partir de los datos recolectados en la primera administración del instrumento. Este ejercicio debe permitir identificar las áreas de oportunidad de la evaluación y con ello planear actividades y metas para atenderlas.

12.3 **Definir las estrategias de la revisión del objeto de medida.** Se deben determinar los procedimientos encaminados a la revisión del objeto de medida del instrumento con el propósito de verificar su vigencia o diagnosticar los aspectos que se deben ajustar, tomando en cuenta lo siguiente:

- **Considerar los análisis del instrumento.** Una fuente importante para detectar la necesidad de actualizar el objeto de medida de la evaluación son los resultados de los análisis de dimensionalidad, de confiabilidad y del funcionamiento diferencial, que permitan suponer un comportamiento errático del instrumento o en alguno de sus componentes o dimensiones. Esta información debe ser el fundamento para orientar la revisión y los posibles cambios para mejorar la precisión de las inferencias que se realizan.
- **Convocar a expertos para el análisis de la información y determinar los ajustes que se requieran.** La revisión cualitativa de contenido debe ser realizada por expertos en el objeto de medida, así como otros especialistas en aspectos relacionados con el desarrollo del instrumento. Se deben documentar los criterios para la selección de los expertos y las evidencias de la idoneidad de los elegidos. La estrategia de revisión debe contemplar integralmente la ficha técnica y el marco teórico o conceptual del instrumento que fundamentan la selección de contenidos y marcan las directrices para la construcción de las especificaciones. Si a partir de la revisión se identifica la necesidad de modificar alguno de los componentes de la estructura, se debe analizar la pertinencia de actualizar la definición del constructo y el marco teórico, además del objeto de medida.

12.4 **Actualizar el objeto de medida**. Si derivado de la revisión del objeto de medida por el Consejo Rector del Instrumento se sugiere que es necesario replantear el instrumento, es obligado pensar en su rediseño, redefinir las especificaciones y renovar el banco de reactivos (College Board, 2015). Es posible que el objeto de medida esté bien definido y únicamente sea necesario realizar algunas precisiones de redacción en alguno de los componentes de la tabla de especificaciones para mejorar su claridad; sin embargo, si se ajusta alguna especificación, se requiere verificar, a través de una validación de expertos, que los reactivos asociados a ella siguen atendiéndola de manera completa.

Cuando una especificación haya perdido vigencia como resultado de la revisión y se recomiende que sea sustituida totalmente o sea omitida de la tabla, los reactivos existentes dejarán de ser útiles y se reconfigurará el instrumento de evaluación.

El resultado de la actualización del objeto de medida se debe presentar al Consejo Rector del Instrumento para su aprobación.

12.4.1 En el caso de instrumentos de respuesta construida, actualizar la rúbrica utilizada con base en la información cuantitativa y cualitativa obtenida de los análisis llevados a cabo para los instrumentos de respuesta construida, así como identificar los aspectos a mejorar en la rúbrica. Los cambios que se propongan deben estar justificados y presentarse al Consejo Rector del Instrumento para su aprobación. Este procedimiento debe documentarse.

12.5 **Establecer los lineamientos para la permanencia de los reactivos o las tareas evaluativas en el banco.** Es necesario documentar los criterios a partir de los cuales se toma la decisión de mantener los reactivos o las tareas evaluativas en el banco.

12.5.1 En el caso de pruebas, se debe determinar cómo se va a controlar la exposición de los reactivos o tareas en las formas (si es el caso), así como la selección, el tratamiento y la sustitución de los reactivos o las tareas ancla.

12.6 **Revisar la pertinencia cualitativa de los reactivos y las tareas evaluativas.** El análisis del banco de reactivos se debe fundamentar en la tabla de especificaciones. Se debe verificar que los reactivos o tareas evaluativas que lo constituyen se apeguen a los lineamientos técnicos de construcción, que corresponden al objeto de medida del instrumento y que evalúan temas vigentes. Asimismo, se deben tomar como referencia los parámetros e indicadores de los reactivos o las tareas evaluativas para hacer la depuración del banco.

12.7 **Documentar los cambios realizados y su justificación**. Debe elaborarse un informe en el que se describan los ajustes realizados derivados de la revisión del instrumento como parte del mantenimiento, así como el sustento para hacerlo. Este documento debe dar cuenta de la evolución del instrumento y de los argumentos en los que se apoyan los cambios realizados.

Evidencias para la verificación documental

- ✓ **Objeto de medida actualizado con la firma del Consejo Rector del Instrumento.** Tabla de especificaciones libre de errores técnicos y de contenido con las firmas de los miembros del Consejo Rector del Instrumento.
- ✓ **Lineamientos para la permanencia de reactivos y tareas en el banco**. Documento que especifique los criterios de permanencia, modificación y baja de los reactivos y las tareas que constituyen el banco.
- ✓ **Banco de reactivos o de tareas evaluativas actualizado**. Se garantiza que los reactivos o las tareas evaluativas que constituyen el banco se apegan a los criterios establecidos en los lineamientos para la permanencia de reactivos y tareas en el banco.
- ✓ ***Currículum Vítae* de los integrantes de los comités académicos de diseño y de actualización de especificaciones, en caso de ser diferentes a los que participaron en la fase de diseño**. Este documento sustenta que las personas que participaron en este paso son expertas en el contenido o en aspectos relacionados con la construcción del instrumento y tienen una trayectoria profesional que les permite determinar las directrices de la evaluación.
- ✓ **Bitácoras que documentan el trabajo de los comités académicos**. Las sesiones de trabajo deben ser respaldadas con las listas de asistencia y los documentos en los que se describan las actividades desarrolladas en cada reunión, así como de los acuerdos tomados acompañados de las firmas correspondientes.
- ✓ **Informe técnico que describa los ajustes realizados en este paso.** Debe referir de manera exhaustiva los ajustes efectuados y su justificación, a partir del diagnóstico realizado.

Paso 13. Revisión y actualización de estadísticos del banco de reactivos o de tareas evaluativas

En este paso se implementa una serie de acciones con el objetivo de incorporar al banco los estadísticos de los reactivos o las tareas evaluativas que sirvan como referentes para valorar la pertinencia de incorporarlos a las formas del instrumento que se aplicarán posteriormente. Una vez que se determina si un reactivo es apto o no para ser utilizado, se le debe asignar una etiqueta que permita su clasificación. Esta información sirve para conocer con certeza la cantidad de reactivos o tareas que pueden ser considerados para ensamblar las formas del instrumento con base en la tabla de especificaciones, lo cual a su vez permite identificar la necesidad de desarrollar más reactivos o tareas evaluativas.

13.1 **Construir un protocolo para establecer el procedimiento de determinación de los parámetros de referencia.** Según las características de administración del instrumento, es necesario que se establezca la metodología para calcular los estadísticos de los reactivos y las tareas que sirvan para dictaminar su pertinencia en cierto periodo.

13.2 **Incorporar los parámetros de referencia al banco de reactivos o tareas.** Se debe concentrar en el banco el historial de estadísticos que los reactivos van acumulando a lo largo de su vida útil y distinguir cuáles se emplearán para el ensamble de las formas.

13.3 **Dictaminar reactivos o tareas evaluativas según su pertinencia estadística.** Todos los reactivos o tareas que están almacenados en el banco deben ser dictaminados de acuerdo con sus estadísticos, a fin de determinar con precisión cuáles podrán considerarse para el ensamble de formas futuras y cuáles no.

13.4 **Vigilar el comportamiento histórico de los reactivos o tareas.** A partir del análisis de los estadísticos que los reactivos o tareas han obtenido a lo largo del tiempo, se debe determinar si se mantienen en uso o es necesario dejar de utilizarlos de manera temporal o definitiva, por ejemplo: si el tema evaluado pierde vigencia, es probable que los parámetros cambien, por lo que se debe analizar colegiadamente la conveniencia de dejar de utilizarlos.

Evidencias para la verificación documental

- ✓ **Protocolo para establecer los parámetros e indicadores de referencia.** Establece el procedimiento para obtener los estadísticos de los reactivos y las tareas evaluativas considerando un largo periodo, así como las políticas de su gestión en el banco.
- ✓ **Banco con parámetros y dictámenes actualizados**. Todos los reactivos y tareas evaluativas que se han incorporado en una forma del instrumento deben tener indicadores estadísticos en el banco, así como su dictamen correspondiente según los estándares establecidos en el protocolo de análisis de la métrica empleada.

Paso 14. Plan de mejora

El plan de mejora consiste en organizar formalmente las metas de producción y las actividades que se efectuarán para sustentar la operación del instrumento. Para realizarlo, es necesario retomar los resultados de la implementación de la actualización del objeto de medida y del banco de reactivos o de tareas evaluativas, así como la revisión y actualización de estadísticos en el banco de reactivos. El plan debe considerar todas y cada una de las actividades que en este documento se describen para el desarrollo del instrumento.

14.1 **Establecer un plan de trabajo para el mantenimiento del instrumento.** Tomando como referencia la información obtenida en los pasos 12 y 13, es obligado establecer metas de trabajo que permitan el funcionamiento adecuado del instrumento.

14.2 **Definir en el plan de mejora las tareas para actualizar el objeto de medida**. El plan debe contemplar las sesiones de trabajo con los comités académicos para el perfeccionamiento del objeto de medida, además de establecer claramente el objetivo de cada sesión.

14.3 **Considerar la construcción de reactivos o de tareas evaluativas para robustecer el banco.** Es necesario que al definir las metas de trabajo se tome en cuenta lo siguiente:

- Inventario del banco de reactivos o de tareas. Según la tabla de especificaciones, se debe determinar la distribución de reactivos o tareas evaluativas, es decir, cuántos aún están en alguna etapa de construcción, cuántos no han sido piloteados y cuántos cuentan con parámetros adecuados.
- Estimación de merma. Calcular la proporción de reactivos o de tareas que no alcanzaron estadísticos adecuados en el piloteo, así como los procesos de revisión y validación, a partir de lo experimentado en el periodo de construcción.

Con base en la información anterior y del tiempo que se dispone, se debe determinar el número de reactivos o de tareas evaluativas a elaborar, revisar, validar y pilotear, considerando la estructura del instrumento y las especificaciones de reactivos.

Evidencias para la verificación documental
✓ **Planeación y cronograma de actividades para el mantenimiento del instrumento**. Establece los pasos que se realizarán, las sesiones de trabajo, su duración y, si corresponde, las metas que se perseguirán para la mejora del instrumento y el robustecimiento del banco.

Paso 15. Informe técnico

El informe técnico del instrumento es un reporte minucioso y detallado de las evidencias que sustentan la validez de las inferencias que se hacen con los resultados de la evaluación. Documenta, de manera sistemática, la implementación de los pasos seguidos en el diseño y la construcción del instrumento; así como lo correspondiente a la administración, análisis y resultados del mismo. Adicionalmente, presenta una serie de recomendaciones para mejorar el instrumento y el proceso de evaluación. Este informe tiene como objetivo organizar y poner a disposición de las personas interesadas toda la información importante para referencias futuras (AERA, APA y NCME, 2014).

15.1 **Describir los antecedentes y los alcances del instrumento.** Se debe dar cuenta del origen y desarrollo del instrumento de manera sintética y describir las razones de su creación; el propósito y alcance del mismo; su desarrollo histórico y evolución; cobertura, población objetivo y participantes; también debe dejar asentado si se han presentado cambios en el objeto de medida y cuándo, así como las razones por las que se dieron dichos cambios.

15.2 **Describir el diseño y la construcción del instrumento.** Debe contener la delimitación del objeto de medida, es decir, describir teóricamente el constructo medido y los contenidos del instrumento e incorporar la descripción de los procesos de elaboración, revisión y validación de reactivos o de tareas, sin dejar de mencionar las estrategias implementadas para garantizar que los reactivos o las tareas evaluativas midan lo que deben medir.

15.3 **Incorporar las funciones desarrolladas por los cuerpos colegiados**. El informe técnico debe dar cuenta, de manera sucinta, del trabajo de los cuerpos colegiados y las actividades en las que participaron dentro del desarrollo del instrumento.

15.4 **Referir la administración del instrumento.** Se debe incluir toda la información sobre la administración del instrumento: periodos, condiciones, requisitos, instrucciones, entre otros.

15.5 **Enunciar las características métricas de los reactivos o tareas evaluativas**. De acuerdo con el modelo estadístico empleado para analizar la pertinencia de los reactivos o las tareas, se deben incorporar en el informe técnico las características de aquellos que forman parte del instrumento y mencionar las reglas de diseño.

15.6 **Incluir información detallada sobre el proceso de puntuación.** Se debe explicar y fundamentar el algoritmo que se emplea para obtener los resultados.

15.6.1 En el caso de pruebas, si el instrumento se interpreta con referencia a un criterio, incluir los estándares de desempeño (descriptores y puntos de corte).

15.6.2 En el caso de pruebas, si es el caso, se debe declarar la estrategia de equiparación de las formas del instrumento, es decir, el método empleado para hacer comparables los resultados de una forma a otra.

15.7 **Presentar las evidencias que sustentan la validez de las inferencias que se realizan, a partir de los resultados de la evaluación**. El informe técnico debe describir las evidencias de la validez de contenido, de criterio y de constructo del instrumento. El contenido medular del documento es el encargado de demostrar empíricamente el grado de la veracidad de las deducciones que se realizan.

No es suficiente mostrar que el instrumento se realizó a partir de la definición del objeto de medida. De acuerdo con AERA, APA y NCME (2014), ETS (2007), JCSEE (2011), AERA, APA y NCME (2014) y Lane, Raymond y Haladyna (2016), se deben realizar estudios de diversa índole:

- análisis factorial confirmatorio. Se comprueba que las variables latentes medidas a través del instrumento se agrupan en las dimensiones definidas de manera teórica;
- relación de la medición con otras variables (convergente y discriminante). Se esperan correlaciones más fuertes con variables con las que debería estar más relacionada, si el instrumento en cuestión está midiendo lo que se propone medir. Las correlaciones fuertes entre medidas teóricamente relacionadas se conocen como evidencia convergente de validez; las correlaciones débiles entre medidas sin relación teórica son evidencia discriminante. Los estudios de validez convergente pueden dar cuenta del poder predictivo del instrumento al demostrarse que los puntajes tienen una fuerte relación con el desempeño futuro de las personas que contestaron el instrumento;
- análisis de funcionamiento diferencial del instrumento y de los reactivos o las tareas. Se espera que el instrumento funcione de manera homogénea en todos los subgrupos que componen la población objetivo.

15.8 **Dar cuenta de los indicadores de confiabilidad.** Presentar los resultados de consistencia interna y de error en la medición. Cada método de cuantificación del error de medida de las puntuaciones debe describirse y expresarse en términos estadísticos congruentes con el método empleado (Wilson, 2005).

15.9 **Considerar el error de muestreo al momento de declarar las inferencias que se pueden efectuar**. Antes de divulgar las conclusiones de una evaluación, es necesario hacer un análisis de los procedimientos utilizados para la recopilación de datos, la representatividad de la muestra en la cual se basan los análisis, las condiciones en que se recogió la información, los resultados de la recopilación de datos (incluyendo los resultados de los subgrupos estudiados de la población), las correcciones o ajustes realizados en las estadísticas reportadas y el grado de precisión de los valores estimados (ETS, 2015; Koretz, 2010).

15.10 **Incluir todas las evidencias posibles para sustentar la calidad métrica del instrumento.** Presentar el sustento empírico con todos los análisis estadísticos posibles. Ninguna evidencia por sí sola es suficiente para establecer que las conclusiones de un instrumento son válidas, aunque una sola de ellas puede demostrar que no lo son (ETS, 2015; Kane, 2013; Koretz, 2010; Lissitz, 2009).

15.11 **Aportar argumentos que permitan interpretar y emplear adecuadamente los resultados**. Elaborar conclusiones que correspondan con las preguntas que guiaron el desarrollo del instrumento y que reflejen con fidelidad los procedimientos empleados y los resultados de la evaluación. Advertir que se debe tener cuidado de no hacer interpretaciones equivocadas de los reportes, incluso se pueden ejemplificar si se considera pertinente. Las inferencias que se pueden realizar se dividen en tres grandes rubros: el puntaje obtenido, la generalización y la extrapolación. Esta clasificación permite dimensionar la complejidad de las inferencias, donde la primera categoría incluye las inferencias más sencillas y la última aquellas más complejas y, por consiguiente, las que deben estar mejor sustentadas, al proporcionar una mayor cantidad de evidencias sobre su pertinencia (Kane, 2013).

15.12 **Integrar todas las evidencias documentales de las fases del proceso del instrumento de evaluación.** El informe técnico debe resumir los aspectos sustantivos documentados de cada una de las fases del proceso de desarrollo de un instrumento de evaluación y que han quedado señalados después de cada fase como EVIDENCIAS PARA LA VERIFICACIÓN DOCUMENTAL.

15.13 **Incluir las fortalezas métricas del instrumento y sus áreas de oportunidad.** El informe técnico debe resumir las características del instrumento que demuestran su solidez y eficacia. También se deben reconocer las acciones que se implementarán para mejorar su calidad técnica.

15.14 **Difundir los beneficios de la evaluación de manera objetiva**. Los materiales de promoción del instrumento deben ser precisos; se debe evitar sugerir que el instrumento ofrece más de lo que está documentado y demostrado empíricamente (Muñiz, 2003). Debe quedar claro a los usuarios de la evaluación que los resultados, aunque son muy útiles, tienen alcances limitados (Koretz, 2010).

15.15 **Incluir la información correspondiente a las adaptaciones al instrumento, para atender a personas con necesidades especiales.** Considerar en este documento toda la información que sustente que las formas con adaptaciones corresponden con el constructo a medir y que sus resultados son equiparables. Incluir las características cualitativas y métricas de dichas adaptaciones al instrumento, así como los indicadores de validez y confiabilidad.

15.16 **Actualizar el contenido del informe técnico a partir de la fase de mantenimiento**. Cada que se realice algún ajuste al marco teórico, al objeto de medida o al tipo de inferencias que pueden realizarse a partir de los resultados del instrumento, es necesario modificar el contenido del documento y notificar a los usuarios para que conozcan la actualización.

15.17 **Revalorar la validez de las inferencias de manera recurrente.** Si ha sido necesario realizar ajustes importantes al instrumento como producto de las acciones de mantenimiento o con el propósito de presentar información vigente, se deben reevaluar periódicamente las evidencias que apoyen al cumplimiento del propósito de la evaluación, y que las interpretaciones previstas de los resultados sean acertadas (ETS, 2015).

15.18 **Describir las estrategias implementadas para controlar el sesgo en las diferentes etapas de la construcción del instrumento.** Se debe dar cuenta de las actividades desarrolladas y del tratamiento que se le dio, con asesoría del comité académico correspondiente, para abordar temas relacionados con la atención a la diversidad, sesgo y diseño universal (Johnstone, Altman y Thurlow, 2006).

15.18.1 En el caso de cuestionarios e instrumentos de autoevaluación, se deben declarar los mecanismos y las herramientas que se emplearon para controlar y atenuar la deseabilidad social, así como los estilos o patrones de respuesta que se hayan identificado en las subpoblaciones evaluadas.

Evidencias para la verificación documental

✓ **Informe técnico del instrumento.** Presenta:

- Características de la evaluación: propósito, población objetivo, objeto de medida, alcances y limitaciones de la evaluación, alcance y usos que se pueden dar a los resultados.
- Objeto de medición: definición del constructo y de las dimensiones que lo componen, de la estructura y de la distribución de reactivos.
- Características del instrumento: tipo, modalidad de administración, condiciones de aplicación.
- Algoritmo de puntuación y reportes de resultados.
- Características métricas de las formas adaptadas.
- Descripción del proceso de construcción del instrumento y de las evidencias que sustentan la validez de las inferencias que se realizan.
- Características métricas del instrumento.
- Reglas de diseño.
- Resultados de los estudios de la validez de las inferencias.
- Estrategias para abordar la sensibilidad del instrumento, los materiales complementarios y el desarrollo de la evaluación (sesgo, atención a la diversidad, diseño universal).

Evidencias adicionales en el caso de pruebas

El informe técnico debe presentar los siguientes apartados:

- Características de la rúbrica (instrumentos de respuesta construida).
- Niveles de desempeño y su descripción.
- Estrategia de equiparación de puntajes.

Evidencias adicionales en el caso de cuestionarios

El informe técnico debe presentar un apartado donde se describa la estrategia para el control de la deseabilidad social, así como los estilos o patrones de respuesta.

Referencias

ACT. American College Testing. (2016). Summative Technical Manual, Iowa City, IA: Author.

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

Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. En NCES Conference on the Program for International Student Assessment: What we can learn from PISA. Washington, DC. Recuperado de: https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

College Board. (2015). Test specifications for the redesigned SAT. College Board. Recuperado de: https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

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(2), 166.e7-166.e16.

Council of Europe. (2011). Manual for language test development and examining. United Kingdom: Association of Language Testers in Europe and Council of Europe Language Policy Division. Recuperado de: http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdfDOF.

DOF. Diario Oficial de la Federación. (2015, 4 de mayo). Ley General de Transparencia y Acceso a la Información Pública. Ciudad de México. Autor. Recuperado de: http://www.dof.gob.mx/nota_detalle.php?codigo=5391143&fecha=04/05/2015

DOF. (2014, 30 de abril). Programa Nacional para el Desarrollo y la Inclusión de las Personas con Discapacidad. México: Autor. Recuperado de: http://dof.gob.mx/nota_detalle.php?codigo=5343100&fecha=30/04/2014

DOF. (2011, 30 de mayo). Ley General para la Inclusión de las Personas con Discapacidad. Ciudad de México. Autor. Recuperado de: http://www.educacionespecial.sep.gob.mx/pdf/doctos/1Legislativos/5Ley_General_Inclusion_de_Personas_Discapacidad.pdf

Downing, S.M. (2004). Reliability: On the reproducibility of assessment data. Medical Education, 38(9), 1006-1012.

Downing, S. M. y Haladyna, T. M. (2006). Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates.

ETS. Educational Testing Service. (2007). Test the francais international. User guide. Recuperado de: https://www.ets.org/Media/Tests/TFI/pdf/TFI_User_Guide.pdf

ETS. (2015). ETS Standards for quality and fairness. Recuperado de: https://www.ets.org/s/about/pdf/standards.pdf

Embreston, S. E. y Reise S. P. (2000). Item response theory for psychologist. Mahwah, NJ: Lawrence Erkbaum Associates.

Fidalgo, A. M. y Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. Educational and Psychological Measurement, 68(6), 940-958.

Ganster, D. C., Hennesey, H. W. y Luthans, F. (1983). Social desirability response effects: Three alternative models. Academy of Management Journal, 26(2), 321-331.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items. (3ª ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. y Rodríguez, M. C. (2013). Developing and validating test items. New York, NY: Routledge.

Havercamp, S. M. y Reiss, S. (2003). A comprehensive assessment of human strivings: Test-retest reliability and validity of the Reiss profile. Journal of Personality Assessment, 81(1), 123-132.

He, J. y Van de Vijver, F. (2016). Corrigiendo las diferencias de uso de escala entre países de América Latina, Portugal y España en PISA. RELIEVE, 22(1), art. M9. Recuperado de: http://dx.doi.org/10.7203/relieve.22.1.8282

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

INEE: Instituto Nacional para la Evaluación de la Educación (2016, junio). Pautas editoriales para la construcción de reactivos, Documentos rectores. Ciudad de México: Autor. Recuperado de: http://www.inee.edu.mx/images/stories/2014/Normateca/Pautas_Editoriales_para_la_construcci%C3%B3n_de_reactivos141216.pdf

INEGI. Instituto Nacional de Estadística y Geografía. (2015, 1 de diciembre). Estadísticas a propósito del día internacional de las personas con discapacidad. Aguascalientes: Autor. Recuperado de: http://www.inegi.org.mx/saladeprensa/aproposito/2015/discapacidad0.pdfI

INEGI. (2010, 12 de diciembre). Acuerdo por el que se aprueba la Norma Técnica para la Generación de Estadística Básica. Aguascalientes: Autor. Recuperado de: http://www.inegi.org.mx/est/contenidos/proyectos/aspectosmetodologicos/documentostecnicos/doc/norma_tecnica_para_la_generacion_de_estadistica_basica.pdf

ITC. The ITC guidelines on the security of tests, examinations, and other assessments. (2014). The security of tests, examinations, and other assessments. Final Version. V. 1.0. Recuperado de: https://www.intestcom.org/files/guideline_test_security.pdf

Jonsson, A. y Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review 2, 130–44. doi:10.1016/j.edurev.2007.05.002

JCSEE. Joint Committee on Standards for Educational Evaluation. (2003). The student evaluation standards: How to improve evaluations of students. Corwin Press. A sage publications Company.

JCSEE. (2011). The program evaluation standards (3ª ed.). Thousand Oaks, CA: Corwin Press.

Johnstone, C., Altman, J. y Thurlow, M. (2006). A state guide to the development of universally designed assessments. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

King, G., Murray, C. J. L., Salomon, J. A., y Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review, 98(1), 191-207.

King, W. y Wand, J. (2006). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis, 15(1), 46-66.

Kolen, M. J., y Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices. (3a ed.). New York, NY: Springer.

Koretz, D. (2010). El ABC de la evaluación educativa (Measuring up). México: Ceneval.

Lane, S., Raymond, M. R. y Haladyna, T. M. (eds.). (2016). Handbook of test development (2ª ed.). New York: Routledge.

Lewis, D. M., Patz R. J., Sheinker, A. y Barton, K. (2002). Reconciling standardization and accommodation: Inclusive norms and inclusive reporting using a taxonomy for testing accommodations. En Supporting inclusion in large-scale assessment: Reconciling standard school testing practices and standardized tests. Simposio llevado a cabo en la reunión anual 2002 de la American Educational Research Association. New Orleans, LA.

Lissitz, R. W. (2009). The concept of validity: Revisions, new directions, and applications. Charlotte, NC: Information Age Publishing.

Mellenbergh, G. J. (2011). A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests. The Hague, Netherlands: Eleven International Publishing.

Muñiz, J. (2003). Teoría clásica de los test. Madrid: Ediciones Pirámide.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. Journal of Educational Measurement. 36(3), 217-232. Recuperado de: http://www.jstor.org/stable/1435155

Paulhus, D. L., y Reid, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60(2), 307-317.

Perie, M. (2008). A guide to understanding and developing performance level descriptors. Center for Assessment. Educational Measurement: Issues and Practice, 27(4),15-29.

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

Schmeiser, C. y Welch, C. (2006). Test Development. En R. L. Brennan (ed.), Educational Measurement (pp. 307-354). Washington, DC: American Council on Education. Praeger Series on Higher Education.

Shun-Wen, Ch. (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., y Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. Teaching of Psychology, 36(2), 102-107.

Stemler, S. E. y Tsai, J. (2008). 3 Best practices in interrater reliability three common approaches. En J. Osborne (ed.), Best practices in quantitative methods (pp. 29–49). SAGE Publications.

Thompson, B. (1990). ALPHAMAX: A program that maximizes coefficient alpha by selective item deletion. Educational and Psychological Measurement, 50(3), 585-589.

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., y Altman, J. R. (2005). 2005 State special education outcomes: Steps forward in a decade of change. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

UNICEF [por sus siglas en inglés]. Fondo de las Naciones Unidas para la Infancia. (2012). Cómo diseñar y gestionar evaluaciones centradas en la equidad. New York: Autor. Recuperado de: http://siare.clad.org/fulltext/2241800.pdf

Van de Vijver, F. y He, J. (2014). Report on social desirability, midpoint and extreme responding in TALIS 2013 (OECD Education Working Papers, No. 107). Paris: OECD Publishing. Recuperado de: http://www.oecd-ilibrary.org/docserver/download/5jxswcfwt76h-en.pdf?expires=1489444173&id=id&accname=guest&checksum=92D7BDD84757C9E18B5FBD970C13817E

Van Vaerenbergh, Y. y Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. International Journal of Public Opinion Research, 25, 195-217.

Ward, A. W., y Murray-Ward, M. (1994). An NCME Instructional Module: Guidelines for the development of item banks. Educational measurement: Issues and practice, 13(1), 34-39.

Wilson, M., (2005). Constructing measures. An item response modeling approach. Mahwah, NY: Lawrence Erlbaum Associates.

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 el artículo 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 nueve de febrero de dos mil diecisiete.- Así lo aprobó la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación en la Tercera Sesión Extraordinaria de dos mil diecisiete, celebrada el nueve de febrero de dos mil diecisiete.- Acuerdo número SEJG/03-17/02,R.- La Consejera Presidenta, **Sylvia Irene Schmelkes del Valle**.- Rúbrica.- Los Consejeros: **Eduardo Backhoff Escudero**, **Teresa Bracho González**, **Margarita María Zorrilla Fierro**.- Rúbricas.

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

(R.- 448331)