Directrices para el control de calidad de las puntuaciones de los tests, su análisis y los informes sobre las puntuaciones

Comisión Internacional de Tests

2013

Traducido por: Ana Hernández

Nota de la traductora: Se han actualizado algunas referencias e incluido referencias a las traducciones españolas de los documentos mencionados en algunos apartados

Adopción formal

El Consejo de la Comisión Internacional de Tests adoptó formalmente las directrices en su reunión de julio de 2012 en Amsterdam, Holanda.

Publicación online

Este documento original fue publicado oficialmente online después de la Reunión General de la ITC celebrada en Julio de 2012 en Amsterdam, y desde entonces se puede obtener a través de la página web de la ITC: http://www.intestcom.org.

La traducción al español puede obtenerse a través de la página web del COP: http://www.cop.es

Publicación en papel

Este documento, en inglés, ha sido publicado en la revista *International Journal of Testing* (2014, volumen 14, pp: 195-217)

Por favor, cite este documento así:

International Test Commission (2013). International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. [www.intestcom.org].

AGRADECIMIENTOS

Estas directrices han sido preparadas para el consejo de la ITC por Avi Allalouf. El autor agradece a Marise Born su valiosa ayuda en la ejecución de este proyecto, así como a las distintas personas que han revisado versiones previas del documento por sus valiosas sugerencias a la hora de desarrollar la versión final:

Alvaro Arce-Ferrer, Pearson Vue, USA

James Austin, Ohio State University, USA

Jo-Anne Baird, Oxford University, UK

Giulia Balboni, University of Valle d'Aosta, ITALY

Helen Baron, Independent Consultant, UK

Dave Bartram, SHL, UK

Marise Born, Erasmus University Rotterdam, NETHERLANDS

James Butcher, University of Minnesota, USA

Janet Carlson, Buros Center for Testing, USA

Iain Coyne, University of Nottingham, UK

Kurt Geisinger, University of Nebraska-Lincoln, USA

Ron Hambleton, University of Massachusetts, USA

John Hattie, University of Auckland, NEW ZEALAND

Fred Leong, Michigan State University, USA

Jason Lamprinou, European University, CYPRUS

Tom Oakland, University of Florida, USA

Fred Oswald, Rice University, USA

Christopher Rush, Wireless Generation, Inc., USA

El autor también está en deuda con sus colegas del NITE (National Institute for Testing and Evaluation), especialmente al departamento de Puntuación y Equiparación, donde se han desarrollado muchos de los procedimientos de control de calidad en los que se centran las directrices

RESUMEN

El propósito de las directrices sobre control de calidad (CC) de los tests es aumentar la eficiencia, precisión y veracidad del proceso de asignación de las puntuaciones obtenidas en los tests, su análisis y la elaboración de los informes correspondientes. Estas directrices pueden ser usadas por sí solas o como una extensión de partes concretas de las directrices internacionales de la ITC sobre el uso de los tests (2000) (ver traducción al español en http://www.cop.es/index.php?page=directrices-internacionales).

Las directrices CC se centran especialmente en los procesos de evaluación a gran escala, cuando se crean múltiples formas de un test para ser usadas en fechas determinadas. Sin embargo también pueden resultar útiles en otros tipos de evaluaciones mediante tests (e.g. evaluaciones individuales para orientación de carrera o desarrollo personal), cuando se usan distintos procedimientos de evaluación (e.g., tests de elección múltiple, evaluaciones de desempeño, entrevistas estructuradas y no estructuradas, evaluaciones de actividades grupales) y para prácticamente cualquier situación de evaluación (e.g. con fines educativos, en centros de evaluación en contextos organizacionales, etc.). Mientras que algunas de las directrices CC son específicas y están relacionadas con ciertos tests estandarizados de administración individual o colectiva, algunos aspectos de las directrices tienen una aplicación mucho más amplia (e.g., en evaluaciones clínicas, educativas y organizacionales). Son muchas las profesiones necesitan hacer evaluaciones -con fines médicos y de rehabilitación, forenses, necesidades especiales, relacionadas con el empleo, etc-, y en estos contextos las directrices CC pueden resultar también muy útiles. Además, las directrices son pertinentes para cualquier forma de administración de tests, desde lápiz y papel hasta las cada vez más frecuentes administraciones informatizadas, tanto a través de internet como sin conexión “on-line”.

CONTENIDOS

INTRODUCCION

Objetivos y finalidad

La estandarización y la precisión son aspectos esenciales en el proceso de evaluación mediante tests, desde la construcción del test y su administración, pasando por la obtención de las puntuaciones y el análisis del test, hasta la interpretación de dichas puntuaciones y la elaboración de los informes correspondientes. Las personas implicadas en cualquier fase del proceso tienen la responsabilidad de mantener unos estándares de calidad profesional que permitan justificar el uso de los tests ante las posibles partes interesadas, incluyendo organizaciones, colegios de psicólogos, institutos y universidades, agencias de gobierno y entidades legales. En este sentido, los usuarios de tests deben ser conscientes de los errores que pueden ocurrir en cualquier fase del proceso, y actuar según las directrices establecidas, con el fin de anticipar, prevenir y abordar dichos errores.

Aplicar una plantilla de respuestas errónea, convertir incorrectamente las puntuaciones directas en puntuaciones típicas, equivocarse al obtener o registrar una puntuación, enviar accidentalmente un informe al cliente equivocado, o interpretar incorrectamente las puntuaciones, constituyen ejemplos de errores que no deberían ocurrir. Aunque errar es de humanos, dichos errores deben minimizarse a través de los procedimientos de control de calidad adecuados. En este sentido los profesionales deben conocer dichos procedimientos, ya que son esenciales para el uso correcto y preciso de los tests. Por ello creemos que este documento contribuirá a la mejora continua de la calidad de los tests y su uso, áreas en que la ITC hace esfuerzos por avanzar.

Las directrices de control de calidad (CC) que se presentan más adelante tienen como objetivo incrementar la eficiencia, precisión y corrección del proceso de puntuación de los tests, su análisis y la elaboración de los informes derivados (proceso PAI). Tienen una doble función: pueden usarse por sí solas, como directrices específicas de control de calidad; pero también pueden usarse como una extensión de partes concretas de las directrices internacionales de la ITC para el Uso de los tests (2000) (ver traducción al español en http://www.cop.es/index.php?page=directrices-internacionales). Se recomienda que el lector esté familiarizado con las directrices de la ITC y con los estándares de la AERA, APA y NCME (1999, 2014), además de otros estándares nacionales e internacionales relevantes.

Destinatarios de las directrices

Las directrices CC se centran en las situaciones de evaluación a gran escala, cuando el test constituye principalmente una medida de rendimiento, desempeño o habilidad (por contraposición a preferencias u otras medidas de autoinforme). Así pues son especialmente aplicables a situaciones de evaluación educativa a gran escala, o en evaluaciones relacionadas con el empleo. Sin embargo, muchas de las consideraciones realizadas aquí, podrían aplicarse también a evaluaciones mediante tests que se realicen a menor escala, o mediante otros tipos de pruebas.

Las directrices CC están dirigidas a las personas responsables de las siguientes cuestiones:

- Diseño y construcción de tests
- Administración de tests
- Obtención de puntuaciones
- Análisis de ítems y de las propiedades del test (incluyendo baremación y equiparación de puntuaciones)
- Mantenimiento de la seguridad del test
- Interpretación de las puntuaciones obtenidas mediante el test
- Elaboración de informes y suministro de feedback
- Formación y supervisión de los usuarios de tests
- Diseño de sistemas informáticos y programas para el manejo de los datos obtenidos mediante tests.

Así como

- Responsables políticos (incluyendo legisladores)
- Editoriales de tests

Ampliar el conocimiento sobre el control de calidad resulta esencial para cualquier profesional implicado en el proceso de evaluación. Aunque las directrices CC están principalmente dirigidas a la práctica profesional de los usuarios de tests, también incorporan una serie de buenas prácticas que son relevantes cuando los tests se usan en investigación, tanto de laboratorio como de campo.

Factores contextuales e internacionales

Las directrices CC están dirigidas a una audiencia internacional de profesionales usuarios de tests. Las directrices pueden ayudar a dichos profesionales a desarrollar estándares de calidad locales. Cuando se interpretan las directrices CC a nivel local, o cuando se considera su utilidad práctica en una situación concreta, es necesario tener en cuenta factores contextuales, como las leyes y estándares nacionales, las regulaciones locales existentes, o los contratos específicos entre clientes y vendedores de tests. Por ejemplo, en algunos países hay leyes que protegen la confidencialidad de los datos personales de las personas evaluadas.

Errores que hacen necesarias las directrices

Los errores que ocurren en el proceso PAI pueden tener serias implicaciones en cualquier ámbito de medición –psicológico, educativo, ocupacional y actitudinal. Por ejemplo, si se comete un elevado número de errores al puntuar un test, el significado de dicha puntuación así como su fiabilidad pueden verse seriamente afectados –la fiabilidad disminuirá, así como la validez predictiva. En algunos casos el error podría suponer que una persona con una conducta patológica sea incorrectamente identificada como una persona con una conducta normal. En otros casos los errores podrían impedir que un candidato cualificado accediera a un determinado puesto de trabajo, o

podrían conllevar una incorrecta asignación de estudiantes en determinados cursos académicos Los errores también podrían resultar en una intervención educativa inadecuada, que conllevara asignar a alguien a un programa educativo inapropiado, o en conceder licencias profesionales o certificaciones académicas a personas que carecen de los conocimientos y habilidades requeridos. Los errores pueden consistir en un excesivo retraso a la hora de informar de los resultados de evaluación, lo que, a su vez, podría causar graves problemas para quienes, por dicho retraso, no pudieran inscribirse en una determinada institución educativa. En resumen, los errores pueden conllevar importantes consecuencias perjudiciales para las personas.

Los errores pueden también contribuir a una pérdida de confianza en los tests educativos y psicológicos, y reducir la credibilidad si los errores se hacen públicos. Los errores pueden, en algunos casos, conllevar acciones legales contra las agencias evaluadoras, las instituciones educativas, los profesionales usuarios de tests y los psicómetras, e incluso contra las empresas que buscan contratar empleados cualificados. Los profesionales que implementan un proceso de evaluación mediante tests (como psicólogos, psicómetras, orientadores, etc.) están sujetos a la presión potencial de cuatro fuentes: las organizaciones, las personas evaluadas, las editoriales de tests y los medios de comunicación. Todas ellas esperan que los tests se desarrollen de forma rápida y económica, y que puedan ser usados tan pronto como sea posible. Para mantener los estándares de calidad, resulta imperativo resistir a la presión ejercida por quienes desean que el proceso se acelere, omitiendo algunas de las fases de dicho proceso. Por ejemplo, podría haber una presión extrema cuando una organización está obligada, por contrato, a puntuar, analizar e informar de los resultados del test en un periodo de tiempo breve. También hay una alta probabilidad de cometer errores en procesos que suelen alargarse en el tiempo como son la construcción de un test, su puntuación (especialmente cuando requiere muchas reglas de puntuación), su análisis y la elaboración del informe de resultados, que suponen pasos secuenciales donde cada paso depende del anterior. El uso de los estándares de calidad ayudará a prevenir estos errores. Para ello se debe hacer un seguimiento regular de los estándares de calidad y actualizarlos cuando sea necesario.

Definición de control de calidad

En este documento, el control de calidad se define como un proceso formal sistemático diseñado para garantizar el mantenimiento de los estándares de calidad en las fases de puntuación de los tests, su análisis y el informe de resultados y, consecuentemente, para garantizar que los errores se minimizan y aumenta la confianza en las mediciones realizadas mediante tests.

Ejemplos de otras profesiones

Los procedimientos de control de calidad se aplican en muchas otras profesiones como la ingeniería, la aviación, el desarrollo de software y la medicina. Por ejemplo, en medicina, algunos de los errores que ocurren en los hospitales son consecuencia del almacenamiento inadecuado de las medicinas, de la complejidad de las intervenciones médicas, de la tecnología novedosa empleada, de la comunicación inadecuada, de un

mal trabajo en equipo y de la ausencia de unas normas de seguridad claras. Este ejemplo tiene su analogía en la evaluación mediante tests, donde el foco es el test, y donde errores potenciales similares amenazan el proceso de administración y evaluación.

Estructura de las directrices de control de calidad

Las directrices CC se dividen en dos partes principales:

1) Principios generales –cuestiones generales a considerar y sobre las que se debe llegar a un acuerdo antes de obtener las puntuaciones, analizar el test y realizar el informe.
2) Directrices de trabajo detalladas, paso a paso

Se concluye con las referencias empleadas

Notas finales

Además de las recomendaciones ofrecidas en este documento, es conveniente presentar algunas directrices generales y sugerencias. Cada vez que se presenta una nueva prueba o procedimiento de evaluación, se debería desarrollar una simulación realista del proceso seguido, paso a paso (ver *Texas Education Agency et al*., 2004). De esta forma los nuevos procedimientos podrían ponerse en práctica y evaluarse. Cada simulación ofrecería información que serviría de input para posibles revisiones de los procedimientos de control de calidad. Además el proceso de obtención de puntuaciones, análisis del test y elaboración de informes consta de fases secuenciales, y cada fase requiere la compleción satisfactoria de la fase previa. Por lo tanto, se recomienda elaborar una lista de verificación –*checklist*- basada en las directrices CC, de forma que resulte imposible pasar a una fase determinada sin que las fases previas hayan sido completadas con éxito. Los sistemas de gestión informáticos podrían ser la herramienta ideal para estandarizar, modificar y controlar los procedimientos CC de forma fácil, transparente y efectiva. Sin embargo, a pesar de que las ventajas de estos sistemas informáticos, es necesario contar con una persona competente y con formación investigadora para desarrollar los procedimientos de control de calidad, adaptarlos y evaluarlos.

DIRECTRICES

Alcance de las directrices de control de calidad

Las directrices CC se centran especialmente en los procesos de evaluación a gran escala, cuando se crean múltiples formas de un test para ser usadas en fechas determinadas. Sin embargo también pueden resultar útiles en otros tipos de evaluaciones mediante tests (e.g. evaluaciones individuales para orientación de carrera o desarrollo personal), cuando se usan distintos procedimientos de evaluación (e.g., tests de elección múltiple, evaluaciones de desempeño, entrevistas estructuradas y no estructuradas, evaluaciones de actividades grupales) y para prácticamente cualquier situación de evaluación (e.g. con fines educativos, en centros de evaluación en contextos organizacionales, etc.). Mientras que algunas de las directrices CC son específicas y están relacionadas con ciertos tests estandarizados de administración individual o colectiva, algunos aspectos de las directrices tienen una aplicación mucho más amplia (e.g., en evaluaciones clínicas, educativas y organizacionales). Son muchas las profesiones necesitan hacer evaluaciones -con fines médicos y de rehabilitación, forenses, necesidades especiales, relacionadas con el empleo, etc-, y en estos contextos las directrices CC pueden resultar también muy útiles.

Además, las directrices son pertinentes para cualquier forma de administración de tests, desde lápiz y papel hasta las cada vez más frecuentes administraciones informatizadas, tanto a través de internet como sin conexión "on-line". La construcción del test, la selección del test y su administración no son objeto de atención de las directrices. Sin embargo, la utilidad o el éxito de la aplicación de las directrices CC para la obtención de puntuaciones, análisis del test y la realización de informes es contingente a que el propio test sea adecuado y a que las puntuaciones obtenidas sean fiables y predictivas de resultados bien definidos. La asignación de recursos para la realización de controles de calidad supone una inversión para asegurar una práctica responsable, una adecuada rendición de cuentas y el mantenimiento de la equidad – aspectos importantes en cualquier código ético.

Parte 1: Principios generales

1.1.Verificación de los estándares de control de calidad existentes

1.1.1 Determinar si existen normas de control de calidad de tests en la organización o país. Si fuera necesario, formular procedimientos de control de calidad específicos para un test antes de su administración. Revisar, actualizar y modificar las normas cuando se realicen cambios en el proceso y también periódicamente, como un chequeo rutinario.

1.1.2. Asegurar que existen procedimientos de control de calidad adecuados antes de la administración del test.

1.1.3. Cuando se trate de un nuevo test, considerar la realización de una prueba o una simulación piloto de todo el proceso PAI. Cuando no se haya realizado una prueba

piloto, tratar la primera administración como un ensayo y estar preparado para implementar mejoras antes de las siguientes administraciones del test.

1.1.4. Crear estándares específicos para cada test en caso de que no existan.

1.1.5. Crear estándares específicos para los tests nuevos que se construyan.

1.2.Cuestiones preliminares y acuerdos entre las personas implicadas

Antes de administrar el test, deben establecerse acuerdos sobre los principios básicos del proceso entre los distintos profesionales responsables de la evaluación, incluyendo los responsables de la construcción del test, su administración, su puntuación, la equiparación de puntuaciones, su interpretación, la validación y la elaboración de informes. De hecho, aunque tengan diferentes responsabilidades y roles, el trabajo de todos los profesionales implicados –vendedores, clientes, socios y colaboradores- debe estar coordinado. La comunicación adecuada entre las personas que juegan distintos roles debería mejorar la calidad y el uso de los tests.

1.2.1. Identificar todas las partes implicadas en el proceso de evaluación y establecer acuerdos sobre quien es responsable de cada una de las fases del proceso.

1.2.2. Determinar y manifestar el propósito o propósitos de la prueba (e.g. selección, medición de desempeño, investigación)

1.2.3. Establecer acuerdos sobre el cronograma a seguir en el proceso PAI

1.2.4. Establecer cuáles son los medios de comunicación más adecuados entre las distintas personas o equipos (cuando haya más de un equipo implicado en la realización de la evaluación); por ejemplo, el mejor modo de transferir información de un equipo a otro; por ejemplo, de transmitir descripciones detalladas sobre la estructura del test, la plantilla de corrección, etc., del equipo que ha desarrollado el test al equipo que se ocupa de su análisis.

1.2.5. Establecer cuáles son los medios más adecuados de comunicación con el cliente

1.2.6. Decidir qué métodos se emplearán para transferir los datos recogidos a las personas responsables del proceso PAI, por ejemplo, en los tests de lápiz y papel, los datos obtenidos mediante lector óptico o escáner, y en los tests informatizados, los datos obtenidos electrónicamente

1.2.7. Definir los pesos o ponderaciones que se emplearán para los subtests (en caso de que se usen) y justificar su elección. Es necesario además estar preparado por si fuera necesario modificar las ponderaciones tras recibir los datos, teniendo en cuenta las justificaciones teóricas existentes y la finalidad del test.

1.2.8. Establecer acuerdos sobre las instrucciones de puntuación del test, es decir, sobre la puntuación que se asignará a cada ítem respondido correctamente, y decidir cómo se tratarán las respuestas incorrectas. Es necesario además estar preparado para modificar las instrucciones, si fuera necesario, tras recibir los datos.

1.2.9. Elegir la escala de puntuación y determinar el rango de puntuaciones en la escala.

1.2.10. Decidir el tratamiento que se dará a los datos faltantes (missing) (e.g. ítems que han sido pasados por alto por las personas que responden la prueba, o que erróneamente han saltado una línea al marcar las respuesta, o casos donde un evaluador ha pasado por alto a un participante o lo ha evaluado de una manera no estandarizada, sin posibilidad de repetir la evaluación).

1.2.11. Cuando las puntuaciones obtenidas mediante versiones diferentes del test deban ponerse en la misma escala, definir y describir el modelo de equiparación de las puntuaciones, el diseño, los tamaños muestrales necesarios, y los métodos de equiparación empleados.

1.2.12. Definir y describir el modelo seguido para elaborar las normas, así como el diseño y tamaños muestrales empleados.

1.2.13. Establecer acuerdos sobre el grado de detalle con el que se informará a las personas evaluadas y a las instituciones implicadas sobre las puntuaciones obtenidas, y qué información adicional se aportará sobre la distribución de las puntuaciones.

1.2.14. Determinar qué individuos, organismos o instituciones recibirán los resultados de las pruebas, asegurando el cumplimiento de las leyes de protección de datos.

1.2.15. Determinar si los informes pueden o deben proporcionar otra información personal o no (e.g. si el contenido del test fue modificado, cuántos ítems fueron respondidos, qué adaptaciones fueron realizadas en caso de discapacidad)

1.2.16. Establecer acuerdos sobre la documentación necesaria para cubrir la totalidad del proceso

1.2. 17. Establecer acuerdos sobre los esfuerzos de replicación que se realizarán cuando se trate de procesos críticos como la conversión de puntuaciones directas a escalas transformadas

1.3.Recursos

1.3.1. Confirmar que se dispone de los recursos adecuados (de costes, tiempo y personal) para poder obtener las puntuaciones de forma adecuada y eficiente, para analizar el test e informar de los resultados

1.3.2. Comprobar la disponibilidad de recursos adicionales en caso de que falle alguno de los disponibles; por ejemplo, si el especialista que se ocupa de la equiparación de puntuaciones no pudiera realizarla, prever quien la llevaría a cabo; o si el escáner lector de las hojas de respuesta no funciona, tener acceso a un escáner alternativo.

1.3.3. Estar al tanto de los problemas de calendario que pudieran darse si fuera necesario utilizar los recursos adicionales mencionados. Planificar la posible necesidad de cubrir bajas inesperadas de personal relevante para la evaluación.

1.3.4. Asignar tareas a los miembros adecuados del equipo: ¿Quién se ocupará de puntuar el test, analizarlo e informar de las puntuaciones? ¿Quién se encargará de supervisar el proceso completo? Los profesionales encargados de la evaluación deben asegurar, por ejemplo, que los individuos implicados en cada fase del proceso tienen las competencias necesarias para realizar el trabajo; también deben establecer los requisitos para cada fase y definir el nivel de automatización del proceso.

1.3.5. Establecer los recursos temporales necesarios: elaborar un cronograma para cada fase del proceso PAI. El plazo para concluir el proceso de evaluación e informar de los resultados debe ser realista.

1.3.6. Determinar la necesidad de software, ordenadores y conexiones a red: software comercial y software desarrollado específicamente para el cliente, ordenadores personales y portátiles, servidores, espacio de disco, banda ancha, etc.

1.3.7. Determinar los espacios de trabajo necesarios: ¿se cuenta con un área de trabajo suficientemente amplia (con suficientes salas, mesas, sillas, etc), para todo el personal y participantes en las pruebas?

1.3.8. Determinar los pasos necesarios para mantener la seguridad de los datos electrónicos

1.3.9. Asegurar la disponibilidad del material que sea necesario (e.g. plantillas de corrección, calculadoras)

1.4.Demandas y expectativas de las partes interesadas

Quienes hacen uso de las puntuaciones de las pruebas –personas evaluadas, padres/tutores, profesores/asesores- y quienes dirigen la evaluación (una agencia, si es el caso) tienen necesidades y expectativas concretas sobre los procesos de puntuación y equiparación y sobre el tiempo necesario para elaborar los informes. Estas necesidades y expectativas deben ser razonables y comunicarse entre las partes (a este respecto ver las directrices internacionales de la ITC para el uso de los tests, 2000, Apéndice B –Directrices para el establecimiento de acuerdos entre las partes implicadas en el proceso de evaluación) (ver traducción al español en http://www.cop.es/index.php?page=directrices-internacionales).

1.4.1. Cuando sea apropiado, formular un acuerdo entre las partes implicadas –partes interesadas, vendedores, participantes en las pruebas, clientes y otros – teniendo en cuenta la opinión de los profesionales responsables de obtener las puntuaciones, realizar la equiparación y elaborar los informes. Se ha de tener en cuenta que, en ocasiones, será necesario realizar cambios en el contrato.

1.4.2. Establecer acuerdos sobre quiénes son los responsables últimos de la evaluación y tienen la autoridad para decidir sobre cómo proceder ante los problemas que surjan y cómo resolverlos.

Por ejemplo, cuando en una pregunta de respuesta múltiple no haya una respuesta correcta, cuando un entrevistador sea muy arrogante, o cuando las personas que participan en las pruebas no puedan concentrarse por el ruido ambiental. También cuando una cuestión se haya construido pensando que sólo una de las respuestas es correcta pero uno de los examinados demuestra que una o varias alternativas adicionales son asimismo correctas.

1.4.3. Decidir por anticipado sobre el proceso a seguir cuando se detecte un error después de que las puntuaciones se hayan dado a conocer

1.4.4. Dar a los participantes la oportunidad de cuestionar la adecuación de las respuestas correctas así como sus puntuaciones, y darles la oportunidad de plantear

cuestiones sobre la evaluación y asegurarles que dichas cuestiones serán tenidas en consideración.

1.4.5. Disponer de documentación que justifique la puntuación de cada ítem del test.

1.5. Personal y ambiente de trabajo

Asegurar que las personas responsables de puntuar el test, analizarlo, equiparar las puntuaciones y elaborar los informes son profesionales que tienen las habilidades y conocimientos requeridos en el proceso PAI. Es decir, asegurar que todo el personal implicado tiene las competencias requeridas para desempeñar adecuadamente el trabajo. Cuando haya un grupo de personas involucradas en el proceso de evaluación es importante que trabajen bien juntas. Por ello, cuando se contrate nuevos empleados, es importante considerar la capacidad del nuevo equipo para trabajar juntos satisfactoriamente.

1.5.1. Evitar ejercer presiones poco razonables sobre los individuos para que aceleren su trabajo.

1.5.2. Evitar horarios de trabajo excesivamente largos

1.5.3. Fomentar una forma de trabajo meticulosa, que preste atención a los detalles (especialmente por lo que se refiere a la prevención de errores), pero que, al mismo tiempo, sea relajada. Un ambiente de trabajo relajado en el que, a la vez, se tiene un propósito claro, es el más efectivo para cumplir estándares elevados.

1.5.4. Apoyar al personal proporcionando oportunidades de desarrollo profesional y formación, e incluso oportunidades de crecimiento personal y entrenamiento en habilidades sociales. Por ejemplo, dar la oportunidad de participar en un sistema de evaluación basado en los datos de un año anterior, como preparación al procesamiento de los datos que se obtendrán en la situación de evaluación actual

1.6. Supervisión independiente de los procedimientos de control de calidad

Asignar uno o más profesionales (dependiendo del tamaño y de la complejidad del proyecto) a la supervisión del seguimiento del proceso CC, y asegurar que todas las cuestiones y problemas que surjan, así como los errores, serán registrados. Los supervisores del procedimiento CC deberían operar de forma independiente a las personas encargadas de puntuar las pruebas, analizarlas y elaborar los informes. La supervisión debería llevarse a cabo en colaboración con las distintas partes interesadas, con el objetivo de auditar procesos específicos; por ejemplo supervisar la fiabilidad inter-jueces y comprobar posibles errores en la introducción de datos. Las asociaciones profesionales podrían adoptar un rol activo en este proceso de supervisión.

1.7. Documentación e informe de errores

1.7.1. Todas las partes implicadas en el proceso de evaluación deberían seguir los procedimientos acordados respecto a la documentación de las actividades y de los errores o cuestiones que pudieran surgir.

1.7.2. Establecer acuerdos sobre qué miembros del personal son responsables de cada fase del proceso.

1.7.3. Documentar todas las actividades. Usar hojas de control estandarizadas para mostrar que todos los procesos han sido comprobados

1.7.4. Documentar con detalle todos los fallos y errores (independientemente de que se conozca o no la causa), comenzando con la naturaleza del fallo, quién lo ha detectado y cuándo, cuáles han sido y son sus implicaciones y qué pasos se han seguido/seguirán para abordarlos. Documentar también los casos en que se hayan detectado fallos antes de que estos hayan tenido consecuencias.

1.7.5. Informar adecuadamente y con prontitud a otros profesionales sobre los fallos observados, por ejemplo en reuniones dedicadas a la prevención de errores

1.7.6. Documentar la forma de prevenir fallos o errores en el futuro.

Parte 2: Las directrices detalladas, paso a paso

Las directrices sugieren una serie de pasos a seguir a la hora de asignar las puntuaciones de los tests, analizarlos y elaborar los correspondientes informes. En procesos de evaluación a gran escala, cada fase debería seguirse minuciosamente. Se debería realizar un estudio piloto sobre los procedimientos de puntuación antes de trabajar con los datos reales, para así agilizar el proceso de obtención de resultados posteriormente. Cuando miles de personas vayan a ser evaluadas, estas directrices deberían seguirse explícitamente. Cuando sólo decenas de personas vayan a ser evaluadas, los principios de las directrices deberían implementarse también, pero algunas fases se podrían omitir o simplificar. La razón radica en que algunos de los procedimientos requieren importantes recursos y están basados en modelos que requieren muestras grandes. Estos procedimientos, por tanto, deberían adaptarse para aplicarlos a muestras menores

2.1. Planificación y diseño del informe

Antes de implementar los distintos pasos, debería establecerse un acuerdo sobre el informe, que es el producto final del proceso. Deberían tomarse decisiones sobre de qué informar, con cuánto detalle, a quién, cuándo, etc. No es suficiente con informar a la institución o a la persona evaluada de la puntuación en el test mediante un número o una escala derivada (estaninos, etc.). Es muy importante interpretar la puntuación adecuadamente. De hecho, en las fases de construcción, puntuación y análisis del test, no se debería perder de vista el producto final: la interpretación de las puntuaciones. En este sentido, el objetivo principal o primer paso tácito del proceso de desarrollo de un test, es asegurar que la puntuación otorgada sea comprendida. Por ello, las distintas

cuestiones relacionadas con la interpretación de las puntuaciones deberían ser consideradas de antemano. Se deberían establecer acuerdos entre todas las partes implicadas sobre las puntuaciones a presentar, no sólo la puntuación total sino también las puntuaciones parciales: ¿Debería darse información al respecto? ¿Se utilizarán esas puntuaciones parciales?

2.2. Antecedentes y datos biográficos

Los antecedentes y los datos biográficos de las personas evaluadas pueden resultar muy útiles para lograr algunos de los objetivos del proceso de control de calidad: Verificar la identidad de la persona evaluada, comprender resultados inesperados y establecer grupos de anclaje cuando sea necesario hacer equiparar las puntuaciones del test. Se recomienda seguir los siguientes pasos:

2.2.1. Si el contexto legal lo permite, recoger datos sobre los antecedentes y biografía (edad, género, grupo étnico, educación, puntuaciones obtenidas en otras pruebas, etc), previamente, durante, o tras haber administrado las pruebas, solicitando esta información a la persona evaluada o a la institución correspondiente. Sólo deben solicitarse los datos que sean relevantes, respetando la privacidad de las personas tanto como sea posible.

2.2.2. Si es posible, comprobar los datos biográficos de las personas evaluadas periódica y sistemáticamente; cuando las personas son evaluadas varias veces, se debe prestar atención a posibles inconsistencias de información.

2.2.3. Realizar estudios para determinar si se da la correlación esperada entre la información contextual y las puntuaciones de la prueba, y buscar posibles inconsistencias entre los patrones de respuesta observados y otros datos o informaciones –conjuntos de datos previos, resultados de investigación, etc. Por ejemplo, podría ser que los adultos hayan obtenido mejores resultados que los jóvenes en un determinado test. Si los estudios sobre el tema sugirieran que los jóvenes deberían obtener mejores resultados en dicho test, el proceso de puntuación debería reexaminarse para determinar si ha habido algún fallo.

2.3. Puntuaciones

2.3.1. Obtención y almacenamiento de las respuestas de las personas evaluadas

Todas las hojas de respuesta de las personas evaluadas deberán guardarse en el lugar apropiado y, cuando sea adecuado, almacenarse electrónicamente, normalmente asignando un número de identificación a cada persona. Estos materiales –impresos y electrónicos- se almacenarán por un periodo mínimo y máximo de tiempo establecido siguiendo los estándares de la práctica profesional y los requisitos legales existentes. Esto es aplicable tanto a las hojas de respuesta identificables individualmente, como a los registros electrónicos de las respuestas o de las puntuaciones, y a la información derivada de dichas puntuaciones.

2.3.1.1. Si existen hojas en papel, deben guardarse por el tiempo establecido según las leyes del país, estado o provincia en cuestión, si dichas leyes existen

2.3.1.2. Por lo que se refiere a las versiones electrónicas, se deben utilizar tanto sistemas de suministro de energía ininterrumpidos como baterías auxiliares para los ordenadores, así como cualquier otro medio que reduzca la probabilidad de pérdidas accidentales de datos.

2.3.1.3. Cuando se usen escáneres estos deben comprobarse y calibrarse regularmente

2.3.1.4. Se debe comprobar manualmente y de forma rutinaria los outputs del escáner

2.3.1.5. Comprobar que la base de datos de las personas evaluadas mantienen un sistema riguroso de identificación. Por ejemplo, comprobar si hay casos donde se haya asignado un mismo código de identificación a distintas personas

2.3.1.6. Todos los datos deben estar protegidos y almacenados de forma segura. Siempre que sea posible se debe proteger la información personal, separando dicha información (e.g. nombres) de las puntuaciones. Por ejemplo, manteniendo ficheros separados: uno con datos biográficos y otro con las puntuaciones obtenidas, pudiendo emparejar ambos ficheros a partir de un código de identificación. Todas estas acciones deberían cumplir las leyes existentes sobre privacidad y almacenamiento de datos.

2.3.1.7 Realizar controles que garanticen la corrección de los algoritmos usados para obtener las puntuaciones, así como el uso adecuado de las tablas de conversión y baremos.

2.3.2. Obtención de puntuaciones

Tras procesar las respuestas de la prueba y almacenarlas de forma segura en una base de datos, las respuestas de las personas evaluadas son habitualmente transformadas en puntuaciones directas. En la Teoría Clásica de los Tests (TCT), por ejemplo, cuando hay respuestas correctas e incorrectas, las puntuaciones directas típicamente se corresponden al número de respuestas correctas obtenidas. En ocasiones se aplica una corrección por posibles aciertos al azar, y en ocasiones se da un peso más alto a unos ítems que a otros. En la Teoría de la Respuesta a los Ítems (TRI) la puntuación directa se corresponde con la "habilidad latente" –también conocida como "theta" o "puntuación en el rasgo". Las puntuaciones pueden verse afectadas por muchos tipos de errores, como podría ser la aplicación de una plantilla de corrección incorrecta. A veces los errores dan lugar a puntuaciones extremadamente bajas. Los procedimientos de control de calidad que se mencionan a continuación pueden contribuir a detectar estos errores.

2.3.2.1. Comprobar si la estructura de los datos se ajusta al formato especificado en el registro de datos (e.g. orden de los ítems en el fichero)

2.3.2.2. Aplicar las reglas acordadas para eliminar casos inválidos, recodificar información faltante y manejar casos duplicados

2.3.2.3. Comparar los datos obtenidos en la muestra con el rango de valores que cabría esperar, y comparar los estadísticos descriptivos obtenidos con los

ofrecidos en los baremos del manual del test (si existen). Cabe esperar ciertas diferencias en los estadísticos muestrales debido al error muestral, pero las diferencias de gran magnitud deberían examinarse y revisarse.

2.3.2.4. Revisar las puntuaciones extremas (altas y bajas), tanto individuales como de grupos específicos, y tanto para tests de lápiz y papel como para tests informatizados. Las puntuaciones extremas podrían ser indicio de tres posibles problemas: un error al calcular la puntuación obtenida en la prueba, una acción deshonesta –e.g. copiar-, o un fallo al obtener los datos.

2.3.2.5. Revisar los datos de las personas evaluadas cuando las diferencias entre las puntuaciones de sub-tests correlacionados sean más grandes de lo esperado. Para ello, debe establecerse de antemano qué diferencias se consideran críticas.

2.3.2.6. Analizar los ítems y examinar los estadísticos correspondientes. A menos que se realice este análisis, los errores en la plantilla de corrección para un ítem concreto serán difíciles de detectar (los ítems corregidos erróneamente suelen mostrar una dificultad elevada y discriminación negativa, y podrían tener una correlación negativa con un criterio relevante).

2.3.2.7. Comprobar las tasas de "no respuesta" para cada ítem. Podría darse el caso de que algún ítem no hubiese sido corregido para algunos participantes al omitirse por error.

2.3.2.8. Prestar especial atención a los grupos que hayan podido responder a la prueba en diferentes condiciones, y realizar comprobaciones adicionales sobre estos datos. Por ejemplo, personas que fueron evaluadas en una fecha diferente, con una versión del test diferente o que usaron un método de respuesta distinto.

2.3.2.9. Revisar los estadísticos básicos obtenidos para ciertos grupos de participantes, por ejemplo según el aula de examen, el administrador de la prueba, o los ordenadores que emplean una misma conexión a internet. Por ejemplo, podría darse el caso de que un test específico hubiera sido erróneamente asignado a un aula de examen determinada.

2.3.2.10. Si hay suficientes recursos, se debería dar una muestra aleatoria de hojas de respuesta a un equipo diferente del asignado inicialmente para analizarla y puntuarla. Posteriormente se podrán comparar los resultados de ambos equipos.

2.3.3. Pruebas abiertas de calificación del desempeño, muestras de trabajo, juegos de rol, entrevistas, etc.

Mientras que la asignación de puntuaciones en los tests de elección múltiple es objetiva y altamente precisa (basada en una plantilla de corrección definida), la asignación de puntuaciones de los ítems de respuesta abierta (calificación del desempeño, cuestionarios de respuesta abierta, muestras de trabajo, juegos de rol, etc) tiene normalmente un componente subjetivo. Este sistema de puntuación tiende a ser consecuentemente menos fiable que el de respuesta múltiple porque requiere de los juicios de los evaluadores. Sin embargo, hay distintos medios que pueden resultar útiles a la hora de reducir la subjetividad de los juicios y así aumentar la fiabilidad y la precisión de las puntuaciones asignadas.

2.3.3.1. Cuando el desempeño es evaluado mediante pruebas abiertas, muestras de trabajo, juegos de rol o entrevistas, hay que asegurar que se cuenta con evaluadores formados para ello, que tienen los conocimientos y la experiencia requeridos, y que cuentan con la certificación, formación o titulación apropiadas

2.3.3.2. Las instrucciones para calificar las respuestas abiertas deben ser claras y estar adecuadamente estructuradas. Realizar un pre-test ayudará a construir dichas instrucciones.

2.3.3.3. Usar ejemplos de respuestas que ejemplifiquen distintos rangos de calificaciones para las distintas actividades. Usar una muestra de respuestas para formar a los evaluadores en la asignación de puntuaciones.

2.3.3.4. Exigir que los evaluadores participen en sesiones de formación antes de comenzar el proceso de evaluación. Esta formación ayudará a que los evaluadores se familiaricen con las instrucciones de calificación y a que practiquen con el sistema de puntuación antes de evaluar a los verdaderos participantes.

2.3.3.5. Antes de comenzar las evaluaciones, comprobar que, mediante la formación, los evaluadores han adquirido las competencias requeridas.

2.3.3.6. Intentar emplear al menos dos evaluadores para cada evaluación individual, dependiendo de los costes y recursos disponibles.

2.3.3.7. Cuando sólo se pueda emplear un evaluador (por cuestiones económicas o de otro tipo) usar dos evaluadores para una muestra de participantes (por ejemplo 10% de los casos) con el fin de estimar la fiabilidad de las pruebas, dependiendo de la importancia de las consecuencias del resultado del test, su longitud y otros factores.

2.3.3.8. Si se emplea un sistema informático para calificar los ítems de respuesta abierta, asegurar que las puntuaciones son supervisadas por un evaluador experto. Justificar el uso del sistema de puntuación informático a partir de los estudios realizados, antes de comenzar a aplicarlo.

2.3.3.9. Asegurar que los evaluadores trabajan de forma independiente

2.3.3.10. Aplicar procedimientos estadísticos para evaluar la fiabilidad del proceso de puntuación (calculando medidas de acuerdo inter-jueces y diferencias en las calificaciones intra y entre evaluadores, ajustando la posibilidad de que se den acuerdos por azar)

2.3.3.11. Supervisar periódicamente y en tiempo real la calidad de las puntuaciones para, si fuera necesario, proporcionar feedback a los evaluadores.

2.3.3.12. Si un evaluador no cumple las expectativas (sus puntuaciones son poco fiables o no son suficientemente parecidas a las de otros evaluadores) informarle de ello y realizar actividades de formación adicionales. Si el problema no se resolviera, reemplazar al evaluador por otro.

2.3.3.13. Desarrollar políticas que permitan responder ante grandes discrepancias entre evaluadores. Si las diferencias son pequeñas, las puntuaciones podrían ser promediadas o sumadas para evitar problemas de redondeo. Cuando las discrepancias sean grandes, un evaluador experimentado podría mediar para resolverlas.

2.4. Análisis del test

2.4.1. Analizar los ítems, normalmente en evaluaciones a gran escala mediante ítems de respuesta múltiple y de respuesta abierta

El análisis de ítems proporciona estadísticos básicos que permiten tomar decisiones sobre las características de los ítems y su funcionamiento a la hora de obtener la puntuación total. Se recomienda que se realice un análisis de ítems en cada ocasión y para cada forma del test, a no ser que el número de participantes sea reducido. El análisis de ítems consiste en obtener la dificultad del ítem (o su "aquiescencia", en ítems de personalidad) y discriminación. Bajo la TRI los parámetros de los ítems pueden ser estimados dependiendo del modelo más adecuado. Además, el análisis de ítems se acompaña de estadísticos globales para el test (fiabilidad y/o error típico de medida, media, desviación típica, función de información del test, distribución de las respuestas de los participantes, etc.). Se debería seguir los siguientes procedimientos siempre y cuando el número de participantes sea superior a un número mínimo, dependiendo del modelo usado:

2.4.1.1. Usar programas confiables para el análisis de ítems y asegurar que dichos programas cuentan con una documentación técnica adecuada.

2.4.1.2. Si hay razones para pensar que el programa de análisis de ítems no resulta adecuado o si se está utilizando un programa nuevo, realizar los análisis con dos programas diferentes y comparar los resultados.

2.4.1.3. Realizar el análisis de ítems tras administrar el test o tras acumular datos sobre un test que se administra periódicamente (por ejemplo entre los 3 y los 5 años de su administración). Considerar realizar los análisis sobre datos parciales (antes de que el total de datos esté disponible), para poder detectar errores rápidamente

2.4.1.4. Revisar los resultados del análisis de ítems antes de extraer conclusiones sobre las personas evaluadas.

2.4.1.5. El análisis de ítems permitirá identificar posibles problemas en la plantilla de corrección de un test. Por ejemplo, si hay distractores tan populares que en realidad son una respuesta correcta; o correlaciones negativas entre los ítems, lo que podría indicar que un ítem que debiera haber sido invertido no lo ha sido. Si los resultados para un ítem particular no son satisfactorios, la corrección y el contenido del ítem deberían revisarse.

2.4.1.6. Repetir el análisis de ítems si la plantilla de corrección es modificada o si algunos ítems son eliminados. Actualizar la documentación (e.g. tablas de puntuaciones, especificaciones de equiparación) a lo largo del proceso.

2.4.2. Equiparación/calibración de nuevas formas de tests e ítems.

En ocasiones, la equiparación no es importante porque los participantes sólo compiten con otros que han sido evaluados en el mismo momento usando la misma versión del test. En otras ocasiones, sin equiparación de puntuaciones, no se podrían comparar los resultados de los participantes a los que se han administrado diferentes formas del test en distintos momentos. Para que las puntuaciones obtenidas con versiones diferentes del test estén en una misma escala, las nuevas formas del test deben equipararse a las anteriores. En caso contrario, podrían no ser comparables al presentar características psicométricas diferentes. El resultado de la equiparación es que las puntuaciones obtenidas mediante las distintas formas del test tienen el mismo significado. Esta equiparación puede realizarse antes de administrar el test y/o después de su administración. La equiparación puede realizarse usando datos a nivel de ítem, escala o test. Hay diferentes perspectivas y métodos de equiparación de puntuaciones (lineal, equipercentil, y basada en TRI -usando ítems comunes de anclaje o individuos de anclaje). La equiparación normalmente requiere muestras grandes, dependiendo del método de equiparación y del diseño (ver Kolen y Brennan, 2004; Lamprianou, 2007)

2.4.2.1. Si tras la equiparación se observan resultados difíciles de explicar (e.g. puntuaciones inferiores a las esperadas), confirmar que todas las formas del test fueron administradas en las mismas condiciones estandarizadas. Si las condiciones de administración no estuvieron estandarizadas, intentar estimar el impacto de las diferentes condiciones.

2.4.2.2. Desarrollar rutinas que aseguren que los procedimientos y diseños de equiparación especificados se han realizado correctamente.

2.4.2.3. Explorar el cumplimiento de los supuestos en que se basa el procedimiento de equiparación y/o determinar si diferentes procedimientos basados en diferentes supuestos ofrecen resultados similares. Realizar una comprobación de la estabilidad de los parámetros de los ítems comunes tras la equiparación. Si se usan ítems comunes de anclaje para la equiparación, documentar la lógica seguida cuando se han eliminado algunos de esos ítems comunes y los efectos que esto pudiera tener tanto para las puntuaciones como para el establecimiento de los puntos de corte. Documentar la representatividad del contenido y las características estadísticas del conjunto de ítems comunes, tras examinar los resultados de los ítems. Esta directriz también se aplica al diseño de "personas comunes como anclaje", pero prestando atención a las personas evaluadas.

2.4.2.4. Comparar las puntuaciones obtenidas con las que se anticiparon en función del historial y antecedentes de los las personas evaluadas (ver 3.2.1). Si existen discrepancias comprobar de nuevo las puntuaciones.

2.4.2.5. Realizar comparaciones con evaluaciones pasadas tanto de las puntuaciones como de las proporciones de aprobados. Cuando las evaluaciones a gran escala se han llevado a cabo adecuadamente, las fluctuaciones de año a año son pequeñas. Diferencias demasiado grandes podrían reflejar un problema en la equiparación de las puntuaciones, o un cambio en las características de la población, por ejemplo.

2.4.2.6. Cuando haya diferentes personas encargadas de administrar la prueba (muchos administradores a cargo de un pequeño número de participantes, en

contraste con pocos administradores a cargo de muchos participantes), aplicar herramientas específicas de control de calidad para supervisar la estabilidad de las puntuaciones del test. Algunas de estas herramientas son: gráficos de control Shewhart y gráficos de control de sumas acumuladas (CUSUM), modelos de series temporales, modelos de punto de cambio y herramientas como la minería de datos (ver Von Davier, 2011).

2.4.2.7. Si hay puntos de corte para diferenciar a los evaluados en función de su nivel (aprobado/suspenso, u otros niveles de rendimiento), comprobar las razones de aprobados y suspensos, o de los distintos niveles establecidos.

2.4.2.8. Asegurar la consistencia de los puntos de corte fijados a través de distintos comités o grupos; usar métodos adecuadamente justificados y documentar el proceso. Asimismo, documentar los casos en los que no se haya seguido completamente el proceso estándar prefijado.

2.4.2.9. Si se usa un formato de administración del test distinto al habitual (e.g. administración informatizada en vez de lápiz y papel) es necesario comparar las características del test con el nuevo formato con las del viejo y, en ocasiones, equiparar ambas formas.

2.4.2.10. Para pruebas que vayan a tener importantes implicaciones sobre las vidas de las personas evaluadas, hacer todo lo posible por replicar de forma independiente los resultados de la equiparación.

2.4.3. Cálculo de puntuaciones estandarizadas

En muchas ocasiones, las puntuaciones estandarizadas ayudan a hacer los resultados más comprensibles. En estos casos el punto de partida para calcular las puntuaciones estandarizadas (e.g. estaninos, deciles) son las puntuaciones directas. Para obtener las escalas transformadas se emplean determinados parámetros o tablas de conversión, y posteriormente se informa de las puntuaciones estandarizadas o los percentiles. Las puntuaciones directas (número de respuestas correctas o número de respuestas correctas tras corregir los posibles aciertos por azar) o las puntuaciones theta (para tests basados en la TRI) deben transformarse en la escala específica seleccionada. La conversión se hace mediante la tabla correspondiente o mediante una determinada función (e.g. transformación lineal)

2.4.3.1. Realizar la conversión de las puntuaciones directas adecuadamente para obtener una determinada escala transformada.

2.4.3.2. Comprobar la precisión de la conversión realizada y posibles errores de copiado

2.4.3.3. Comprobar que se ha utilizado la conversión correcta.

2.4.3.4. Verificar que las puntuaciones estandarizadas bajas corresponden a puntuaciones directas bajas, y que las puntuaciones estandarizadas altas corresponden a puntuaciones directas altas

2.4.3.5. En algunos casos deberían aplicarse procedimientos adicionales tras la realización de la conversión (e.g. definir un mínimo y un máximo uniformes para cada una de las puntuaciones reportadas)

2.4.3.6. Comparar las propiedades de nuevas formas del test con las tablas/parámetros de otras formas del test, con el fin de detectar si se dan discrepancias poco esperables

2.4.3.7. Tener en cuenta los cambios que se dan en la escala a lo largo del tiempo

2.4.3.8. Calcular algunas puntuaciones manualmente y comparar los resultados con los generados por el ordenador.

2.4.3.9. Comprobar la relación estadística entre las puntuaciones directas y las estandarizadas usando gráficos de dispersión.

2.4.3.10. Usar dos programas informáticos diferentes para obtener las puntuaciones estandarizadas y compararlas.

2.4.3.11. En el manual técnico del test o mediante documentación adicional, proporcionar una descripción detallada de los procedimientos usados para transformar las puntuaciones directas en estandarizadas. Puesto que la técnica puede ser diferente para diferentes formas del test, el procedimiento debería describirse para cada una de las formas

2.4.4. Comprobaciones de la seguridad de los tests

Si se descubre que una puntuación ha sido obtenida mediante trampas o engaños, esto supone un serio problema que compromete tanto la seguridad y la integridad del test como el sistema de evaluación. Desgraciadamente este problema no puede prevenirse completamente, incluso aunque se pongan medidas para ello. La tentación de hacer trampa es grande, especialmente cuando los resultados de la evaluación conllevan consecuencias importantes. En la continua batalla contra las trampas, se debería contar con el asesoramiento de abogados para revisar los controles de seguridad y confirmar su aplicabilidad. En pruebas educativas nacionales, el fraude puede darse a nivel individual pero también a nivel de clase, escuela, distrito o lugar de trabajo. Podría ocurrir en el lugar que se realiza la prueba, a través de los teléfonos móviles, o a través de internet. En el contexto organizacional, puesto que cada vez es más frecuente que los candidatos para un puesto de trabajo realicen las pruebas desde casa (vía internet), el riesgo de suplantación y manipulación aumenta. Adicionalmente, al realizar controles de seguridad para la detección de posibles fraudes, se podrán detectar problemas en la administración del test, o en la recogida y almacenamiento de datos. Se recomienda tomar las siguientes precauciones:

2.4.4.1. Verificar la identidad de todos los examinados en el momento de entrar a la sala. Cuando realicen la prueba desde casa, usar un carnet identificativo con fotografía para comprobar la identidad o tomar medidas biométricas como la lectura del iris o de las huellas dactilares. También pueden usarse otras técnicas más avanzadas para verificar la identidad de quienes participan a distancia.

2.4.4.2. Es aconsejable usar múltiples formas del test. Cuando sólo se use una forma, las personas que pudieran conocerse (e.g. vecinos, compañeros) no debieran sentarse juntas. Se puede, por ejemplo, sentar a los participantes por orden alfabético.

2.4.4.3. Numerar los asientos y hacer un listado de dónde se sentó cada persona para ayudar a detectar si se ha copiado.

2.4.4.4. Cuando resulte apropiado (e.g. si se sospecha de que se ha copiado), emplear índices estadísticos que ayuden a detectarlo, basados en la similitud de las respuestas de los participantes ubicados en el mismo aula o lugar de la prueba.

2.4.4.5. Utilizar personal entrenado y fiable que supervise la prueba y controlar su trabajo regularmente. Asegurar que dichos supervisores no tienen ningún conflicto de intereses

2.4.4.6. Comprobar la existencia de patrones de respuesta aberrantes o inesperados (e.g. cuando los ítems difíciles son respondidos correctamente y los fáciles se fallan)

2.4.4.7. Obtener una muestra de la escritura de cada participante antes y durante el examen, para ayudar a detectar suplantaciones. Este procedimiento puede obviarse si no hay problemas de identificación.

2.4.4.8. Cuando haya personas que repitan las pruebas (si estas personas pueden ser identificadas), analizar la diferencia de puntuaciones usando una distribución estadística de diferencias razonables entre la puntuación obtenida en la última prueba y las obtenidas en administraciones previas. Diferencias extremas podrían indicar que la persona en cuestión ha sido suplantada por otra o que ha obtenido información sobre los ítems de la prueba antes de su administración. Otra posible explicación radicaría en los efectos de práctica resultantes de realizar la misma prueba o una prueba similar repetidamente.

2.4.4.9. Documentar (legalmente, si fuera necesario) el procedimiento a seguir con las personas sospechosas de hacer trampas en la prueba. Informar a los participantes por anticipado de que se han puesto en marcha procedimientos para combatir el fraude.

2.4.4.10. A veces, los profesores podrían tener interés en mejorar las puntuaciones obtenidas por sus estudiantes en las pruebas estandarizadas. Por esta razón, los profesores no deberían tener acceso a las mismas.

2.4.4.11. Usar armarios que puedan cerrarse con llave y servidores seguros para guardar con seguridad los materiales del test y sus resultados. Asegurar que las personas implicadas en elaborar los ítems del test son de confianza y siguen las normas establecidas para mantener su seguridad. Además la confidencialidad de los ítems del test debe asegurarse de principio a fin, incluyendo el borrador de los ítems del test. Los ítems tienen que transmitirse de forma segura entre los vendedores de las pruebas y sus creadores, y todos los archivos deben mantenerse y procesarse en tarjetas de memoria o en ordenadores de uso independiente, y no en ordenadores o servidores a los que pudieran tener acceso personas no autorizadas.

2.4.4.12. Los ordenadores que se usen para administrar las pruebas deben tener inhabilitadas las opciones que permitan guardar o enviar los materiales de la prueba. Se debe evitar el acceso a internet si este acceso permite el envío de materiales.

2.4.2.13. Para mantener seguros los materiales del test, asegurar que éstos no se fotografían (mediante cámara o teléfono móvil)

2.4.4.14. Para asegurar el trato equitativo de todos los participantes, su anonimato debe garantizarse en todas las fases de la realización de la prueba y su puntuación.

2.5. Elaboración de informes

2.5.1. Informe sobre las puntuaciones

Las puntuaciones son comunicadas tanto a las personas evaluadas como a los usuarios (clientes). Idealmente, los informes sobre las puntuaciones deberían proporcionarse en un formato imprimible. En ocasiones se usa internet como método estándar para informar de los resultados. En cualquier caso la información debe proporcionarse de forma que el significado de las puntaciones quede claro tanto para la persona evaluada como para el cliente.

2.5.5.1. Usar grupos focales, procedimientos de "pensar en voz alta", estudios experimentales" o incluso entrevistas individuales para obtener información que ayude a generar explicaciones comprensivas y útiles de las puntuaciones así como guías interpretativas.

2.5.1.2. Asegurar que quienes reciben las puntuaciones tienen la ayuda necesaria para interpretarlas, y así poder comprenderlas. Aportar evidencia de que los informes son comprensibles para los destinatarios.

2.5.1.3. Los informes generados informáticamente deben ser apropiados para los destinatarios, facilitando una interpretación de las puntuaciones y los datos más técnicos

2.5.1.4. Si es necesario, usar repositorios de datos donde los resultados de pruebas transnacionales, nacionales y provinciales puedan subirse al momento.

2.5.1.5. Clarificar el nivel al que las puntuaciones pueden interpretarse de forma fiable (e.g. cuando haya subescalas con baja fiabilidad). La decisión sobre si se debe presentar información de las puntuaciones obtenidas en los subtests debe basarse en la teoría que subyace a la prueba, el objetivo de la evaluación y las propiedades psicométricas de las puntuaciones de las subescalas.

2.5.1.6. Buscar ayuda de expertos en relaciones públicas cuando los informes sobre las puntuaciones deban presentarse a políticos y medios de comunicación.

2.5.2. Medidas para mantener la seguridad de los informes

2.5.2.1. Tomar precauciones para que los informes individuales no puedan ser falsificados por las personas evaluadas

2.5.2.2. Evitar realizar correcciones manuales en los informes institucionales. Si es necesario cambiar una o más puntuaciones, usar el software apropiado o crear de nuevo el informe.

2.5.2.3. Encriptar los ficheros electrónicos de los informes para guardar y transferir la información.

2.5.2.4. Asegurar que los informes sólo se envían a las personas apropiadas. No enviar informes que incluyan más información de la necesaria. Sería más fácil enviar el informe completo a todos los usuarios de los tests, pero con el fin de salvaguardar la confidencialidad de los participantes, sólo se enviarán los resultados relevantes para él o ella.

2.5.2.5. Informar a las instituciones de que sólo deben usar el informe enviado directamente a la institución, y no una copia del informe enviado a la persona evaluada (que podría haberse falseado). También es recomendable que las instituciones realicen verificaciones rutinarias de los informes institucionales.

2.5.3. Documentación

La documentación que recoge información exhaustiva del proceso de obtención de puntuaciones, incluyendo estadísticos descriptivos clave (media, desviación típica, mediana, rango de puntuaciones, fiabilidad, etc.), y la compara con las puntuaciones obtenidas por otros grupos de participantes, debe estar preparada y completada antes, o poco después, de haber ofrecido los resultados. Una adecuada documentación contribuirá a aumentar la fiabilidad y precisión del proceso. Hacer pública parte de esta información puede ser un método adicional de control del proceso PAI. Es importante:

2.5.3.1. Documentar el proceso completo, paso a paso (informe interno). Debe incluir documentación estandarizada sobre el proceso de obtención de las puntuaciones, incluyendo los estadísticos principales y las comparaciones entre grupos.

2.5.3.2. Asegurar que las posibles nuevas formas de un test son administradas sólo después de haber completado la documentación para una forma anterior.

2.5.3.3. Compilar estadísticos descriptivos, por ejemplo, por lo que se refiere a diferencias de género y año, y permitir que el público general tenga acceso a estos estadísticos. Se debería proporcionar una breve explicación sobre la interpretación de estos estadísticos. Los estadísticos a nivel agregado protegen la anonimidad de los participantes individuales.

REFERENCIAS

AERA/APA/NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

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

Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. *Educational Measurement: Issues and Practice, 26*: 36-43.

Bartram, D., y Hambleton, R.K. (Eds.) (2006) .*Computer-Based Testing and the Internet.* West Sussex: John Wiley & Sons.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

ITC (2001). International Guidelines on Test Use. *International Journal of Testing, 1:* 95-114.

ITC (2006). International Guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6:* 143-172.

Kolen, M. J., y Brennan, R. L. (2004). *Test equating, linking and scaling: Methods and practices.* New York: Springer.

Lamprianou, I. (2007). *Comparability methods and public distrust: an international perspective.* En Newton, P., Baird J., Goldstein, H., Patric, H., y Tymms, P. (Eds.) Techniques for monitoring the comparability of examination standards. Qualifications and Curriculum Authority, London.

Nichols, S. L. y Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*, Educational Policy Studies Laboratory, College of Education, Arizona State University.

Rhoades, K., y Madaus, G. (2003). *Errors in standardized tests: A systemic problem.* (NBETPP Monograph). Boston, MA: Boston College, Lynch School of Education.

Texas Education Agency, Pearson Educational Measurement, Harcourt Educational Measurement & Beta, Inc. (2004) Capítulo 9: *Quality control procedures. Texas Student Assessment Program*. Technical Digest (2003-2004) http://www.tea.state.tx.us/student.assessment/resources/techdig04/

Toch, T. (2006). *Margins of error: The testing industry in the No Child Left Behind era.* Washington: Education Sector Report.

Von Davier, A. (2011) *Statistical Models for Test Equating, Scaling, and Linking.* Springer

Wild, C. L., y Rawasmany, R. (Eds.) (2007). *Improving testing: Applying process tools and techniques to assure quality.* Mahwah, NJ: Erlbaum.

Zapf, D. y Reason, J. (1994). Introduction: Human Errors and Error Handling. *Applied Psychology: An International Review, 43*: 427-432.