PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-26

Pontificia Universidad Católica de Chile
Facultad de Educación

Alentando los distintos actores a evaluar informes de resultados de pruebas estandarizadas: desafíos y propuestas

Empowering End Users to Evaluate Score Reports: Current Challenges and Proposed Solution

Fernanda Gándara y Francis Rick
The University of Massachusetts, Amherst, United States

Resumen

El informe de resultados es uno de los pasos más críticos para el uso válido de pruebas estandarizadas, aun así, los usuarios de estas no están necesariamente conscientes de su relevancia. Es importante notar que los destinatarios no siempre poseen el conocimiento o los instrumentos para evaluar la calidad de un informe de resultados, ya que no hay una guía explícita en la literatura que los alienta a hacerlo. Los informes de resultados son algunas veces acompañados con guías de interpretación, que permite a los participantes interpretar los datos, pero que no les permite desarrollar una evaluación independiente y crítica de las evaluaciones recibidas en los informes. En este estudio analizamos la forma de evaluación proporcionada en el modelo de Hambleton y Zenisky (HZ) para el desarrollo de informes de resultados para entender hasta qué punto el formulario resulta claro, útil y valioso para evaluaciones dentro de distintos contextos. Se utilizó un grupo focal a pequeña escala donde se documentó los principales problemas con el formulario HZ para la evaluación de informes de resultados. Se pudo identificar modificaciones y directrices apropiadas para el formulario, para conducir, finalmente, estudios subsecuentes que permitan desarrollar una evaluación por parte de los usuarios para examinar la calidad de los informes de evaluación.

Palabras clave: informes de evaluación, evaluación, validación, medición educacional

Correspondencia a:
Fernanda Gándara
Email: fergandara@gmail.com

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.11

Abstract

Score reporting is one of the most critical steps to the valid use of test scores, yet end users are not necessarily aware of their relevance. Importantly, end users do not always possess the knowledge or instruments to evaluate the quality of a score report, as there is no explicit guidance in the literature that empowers them to do so. Score reports are sometimes accompanied with interpretive guides, that allows stakeholders to make better sense of the data, but that do not enable end users to develop independent and critical evaluations of the reports that they receive. To that end, in this study we analyze the evaluation form provided in the Hambleton and Zenisky (HZ) model for developing score reports, to understand to what extent this form is clear, useful, and meaningful to evaluate these across different contexts. Using a small scale focus group, we were able to document the main problems with the HZ form to evaluate score reports. We were also able to identify appropriate directions and modifications to the form, to ultimately conduct subsequent studies that allow us to develop an end-user evaluation for to assess the quality of score reports.

Keywords: score reporting, testing, validity, educational measurement

El informar resultados de evaluación es uno de los aspectos más desafiantes del desarrollo de tests (Zenisky & Hambleton, 2012). Algunos informes transmiten un mensaje, normalmente entre las agencias de evaluación y grupos específicos de actores interesados. Los distintos actores interesados (o los usuarios relevantes de datos de evaluación) tienen sus propias preferencias con respecto a cómo recibir datos de evaluación (Jaeger, 2003; Zwick, Zapata-Rivera, & Hegarty, 2014); por lo tanto, es un desafío central crear informes que cumplan los requisitos de distintos grupos. Otro desafío tiene relación con el hecho de que el contenido de estos informes puede ser de distintos tipos: el mensaje puede ser de carácter sumativo, diagnóstico o normativo (Hambleton & Zenisky, 2013). Al preparar cada tipo de informe (sumativo, diagnóstico o normativo), las agencias deben tomar decisiones con respecto a qué tipo de puntuaciones mostrar, qué información adicional incluir y qué formato o diseño emplear. Además, estas decisiones deben estar de acuerdo con las interpretaciones y usos esperados de las puntuaciones de cada test. Asimismo, los informes de cada test pueden entregarse en medios distintos. Tradicionalmente, los informes de resultados en papel eran los más comunes, aunque hoy en día es cada vez más común que los programas de evaluación entreguen sus informes a través de Internet (Zenisky & Hambleton, 2012). Los informes en línea traen consigo desafíos adicionales. Por ejemplo, al confeccionar informes en línea, se debe considerar hasta qué punto se permitirá la interactividad para el usuario y cómo se entregarán materiales de interpretación (Zenisky & Hambleton, 2012). Debido al gran número de decisiones involucradas y las variadas formas posibles de enfrentarlas, las prácticas relacionadas con los informes de resultados, tanto en papel como en línea, difieren de modo sustantivo entre agencias (Goodman & Hambleton, 2004; Knupp & Ansley, 2008; Faulkner-Bond, Shin, Wang, & Zenisky, 2013).

Muchos profesionales de los tests dejan la confección de informes para el final del proceso de desarrollo de los tests y no prestan suficiente atención al enfoque utilizado (Zenisky, Hambleton, & Sireci, 2009). Por su parte, quienes estudian cuidadosamente el proceso observan que es necesario tomar muchas decisiones y terminan por considerar la preparación de informes de resultados como una tarea abrumadora. Afortunadamente, la literatura psicométrica entrega orientación a través de modelos para desarrollar informes de resultados, estándares generados por asociaciones profesionales y listas de buenas prácticas y/o prácticas que deben evitarse al desarrollar informes de resultados. En primer lugar, la literatura entrega modelos para desarrollar informes de resultados efectivos, como los de Zapata-Rivera (2011) y Hambleton y Zenisky (2013). Estos modelos tienen algunas características esenciales en común. Ambos se basan en la investigación. Ambos hacen hincapié en la importancia de apuntar a públicos específicos y entregan recomendaciones similares con respecto a cómo satisfacer sus necesidades. Ambos modelos se basan en la idea de una línea de producción donde los informes primero se crean, generándose luego prototipos de uso interno y externo que finalmente se ajustan para producir informes de resultados adecuados y útiles (Hambleton & Zenisky, 2013; Zapata-Rivera, 2011; Zenisky & Hambleton, 2015). Con respecto a sus diferencias, sólo Hambleton y Zenisky (2013) abordan los problemas relativos a monitorear y corregir/rediseñar los informes de resultados. Ambos modelos son similares, pero el de Hambleton y Zenisky es más exhaustivo y entrega más detalles para guiar el trabajo de los desarrolladores de tests.

En segundo lugar, las asociaciones profesionales abocadas a los tests publican estándares para estimular el desarrollo y uso correcto de los mismos. Los más ampliamente conocidos son los Estándares para pruebas educativas y psicológicas [*Standards for Educational and Psychological Testing*] (de aquí en adelante *Estándares*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 2014). Los *Estándares* entregan al menos 13 directrices que afectan directamente el proceso de informar resultados. En general, estos estándares pueden clasificarse en tres categorías: (a) los referidos al proceso de desarrollar informes de resultados, (b) los referidos a las responsabilidades de los programas de evaluación con respecto a la interpretación de los informes y (c) los referidos al contenido de los informes de resultados. La Tabla 1 entrega más detalles.

Tabla 1

Estándares referidos a informes de resultados extraídos de los Estándares para pruebas educativas y psicológicas

Tema	ID	Contenido
Proceso	6.0	Para apoyar las interpretaciones útiles de las puntuaciones, los instrumentos de evaluación deben tener procedimientos . . definidos para la confección de informes. Las personas a cargo de la generación de . . informes . . deben tener capacitación y apoyo suficientes que les ayuden a seguir los procedimientos establecidos. La adherencia a los procedimientos establecidos debe monitorearse y cualquier error importante debe documentarse y corregirse si es posible.
	6.13	Cuando se encuentre un error relevante en las puntuaciones de un test o en otra información importante emitida por una organización de evaluación o de otro tipo, dicha información debe distribuirse junto con un informe de resultados corregido lo antes posible a todos los receptores conocidos, quienes de otro modo podrían usar los resultados erróneos para informar su toma de decisiones. El informe corregido debe marcarse como tal. Debe documentarse lo que se hizo para corregir los informes. Las razones para corregir el informe deben explicarse claramente a los receptores.
	9.16	A menos que las circunstancias claramente requieran que los resultados se mantengan en reserva, el usuario de un test está obligado a entregar un informe oportuno de los resultados a quien rinde el test y a otras personas con derecho a recibir esta información.
	9.20	En situaciones en las que los resultados de los tests se hacen públicos, los usuarios deben formular y compartir la política definida con respecto a la publicación de los resultados (por ejemplo, puntualidad, nivel de detalle, etc.) y aplicar dicha política de forma sistemática a través del tiempo.
Responsabilidades de los programas de evaluación con respecto a la interpretación	6.10	Cuando se libera la información sobre los resultados de un test, las personas a cargo de los programas de evaluación deben entregar interpretaciones que se adecuen a su público objetivo. Las interpretaciones deberían describir, en un lenguaje simple, los aspectos que cubre el test, lo que representan los resultados, la precisión/confiabilidad de los resultados y cómo se espera que se usen dichos resultados.
	12.15	Las personas a cargo de los programas de evaluación educativa deben tomar las medidas necesarias para verificar que los individuos que interpreten los resultados para tomar decisiones en contextos escolares estén calificados para hacerlo o que, en su defecto, puedan recibir ayuda de personas con dichos conocimientos.
Contenido	2.4	Cuando una interpretación de los resultados de un test hace hincapié en las diferencias entre dos puntuaciones de un individuo o entre dos promedios de un grupo, deben entregarse los datos de confiabilidad/precisión de dichas diferencias, incluyendo el error estándar.
	3.17	Cuando se informan públicamente las puntuaciones globales de subgrupos relevantes, … los usuarios de los tests tienen la responsabilidad de entregar evidencia sobre comparabilidad y de advertir sobre la existencia de investigaciones teóricas fidedignas que indiquen que los resultados del test podrían no tener un significado comparable entre dichos subgrupos. Cuando en un informe se incluyen interpretaciones de protocolos de respuesta a tests o de desempeño en tests, deben presentarse las fuentes, justificaciones y bases empíricas de dichas interpretaciones; además, deben describirse sus limitaciones.
	6.12	Cuando se obtiene información grupal combinando los resultados de tests parciales rendidos por individuos, deben entregarse los indicadores relacionados con validez y confiabilidad/precisión de acuerdo al nivel de agrupación en el que se informen los resultados. No deben informarse resultados individuales sin evidencia apropiada que apoye las interpretaciones para los usos esperados.

12.17	En contextos educativos, los informes que indiquen diferencias en las puntuaciones alcanzadas por un grupo en un test deben acompañarse con información contextual relevante siempre que sea posible, de modo de permitir la interpretación significativa de las diferencias. Cuando no esté disponible ninguna información contextual apropiada, los usuarios deben recibir advertencias sobre posibles interpretaciones erróneas.
12.18	En contextos educativos, los informes de resultados deben ir acompañados de orientaciones claras sobre cómo interpretarlos, incluyendo el grado de error de medición relacionado con cada puntuación o nivel de clasificación, además de información suplementaria sobre resúmenes de puntuaciones grupales. Además, en los informes de resultados deben incluirse las fechas de administración de los tests junto con estudios de normatización relevantes.
12.19	En contextos educativos, cuando los informes de resultados incluyen recomendaciones para intervenciones instructivas o están asociados a planes o materiales de instrucción recomendados, deben entregarse justificaciones y evidencias para apoyar dichas recomendaciones.

La Comisión Internacional de Tests (International Test Commission, ITC) desarrolló otro grupo de estándares para la confección de informes de resultados. Entre los años 2000 y 2012, la ITC desarrolló y publicó cuatro series de lineamientos para potenciar las prácticas responsables con respecto a los tests. La Guía de control de calidad para la asignación de puntuaciones, análisis de tests y confección de informes de resultado [*Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores*] (ITC, 2012) busca apoyar el funcionamiento de evaluaciones a gran escala a nivel mundial. La Guía entrega lineamientos (no estándares) y debiera adaptarse para desarrollar estándares utilizables en contextos locales. Al menos 15 de las directrices incluidas en la Guía se refieren directamente al tema de la confección de informes. En general, las directrices pueden agruparse en cuatro categorías: (a) las relacionadas con el proceso, (b) las relacionadas con la interpretación, (c) las relacionadas con el contenido y (d) las relacionadas con la seguridad de los informes de resultados. La Tabla 2 entrega más detalles.

Tabla 2

Estándares referidos a informes de resultados extraídos de la Guía ITC de control de calidad para la asignación de puntuaciones, análisis de tests y confección de informes de resultados

Tema	ID	Contenido
Proceso	1.2.1	Identificar a todos los actores interesados en el proceso de evaluación y llegar a un acuerdo con respecto a quién está a cargo de tomar decisiones sobre las diferentes partes del proceso de evaluación.
	1.2.2	Determinar y explicitar el propósito o los propósitos del uso del test (por ejemplo, selección, medición de logro, investigación).
	1.2.14.	Determinar específicamente qué individuos, organismos o instituciones deben recibir los resultados del test para así cumplir las leyes sobre privacidad de datos.
	1.3.1.	Confirmar que existan recursos adecuados (costo, tiempo y personal) para asignar puntuaciones, analizar el test y confeccionar informes de resultados.
	1.4.3	Decidir con antelación el proceso para tratar casos en que se descubra un error después de informar los resultados.
Interpretación	2.5.1.1.	Usar grupos focales de personas que rindan el test, "procedimientos de pensamiento en voz alta", "estudios experimentales" o incluso "entrevistas personales" para obtener información que contribuya al desarrollo de explicaciones comprensibles e instructivas del informe de resultados y a la creación de guías interpretativas.

	2.5.1.2.	Asegurarse de que toda persona que reciba las puntuaciones tenga una orientación adecuada para interpretarlas, de modo que se comprendan adecuadamente los resultados. Se debe apoyar esta información con evidencia de que los informes permiten que los usuarios hagan interpretaciones lógicas.
Contenido	1.2.13.	Llegar a un acuerdo sobre el nivel de detalle con el que deben informarse las puntuaciones a las personas que rinden el test y a las instituciones involucradas, así como también sobre qué información adicional debe entregarse sobre distribuciones de puntuaciones y uso de las mismas.
	1.2.15.	Determinar si los informes pueden o deben incluir otra información personal (por ejemplo, modificaciones al contenido del test, el número de ítems completados o las facilidades que se ofrecieron a personas con alguna discapacidad).
	2.5.1.5.	Aclarar hasta qué nivel los resultados son confiables (por ejemplo, en casos en que las subpuntuaciones tengan una confiabilidad demasiado baja para usarse en la toma de decisiones de alto riesgo). La decisión de informar o no informar las puntuaciones de los subtests también debe basarse en (a) la teoría del test y (b) el objetivo del test y las propiedades psicométricas de las puntuaciones de los subtests.
Seguridad	2.5.2.1.	Tomar medidas para asegurar que los informes de resultados individuales no puedan ser falsificados por las personas que rindan el test.
	2.5.2.2.	En lo posible, no editar el informe para instituciones: hacerlo podría causar problemas serios. Si es necesario cambiar una o más puntuaciones, se debe usar el software correspondiente o volver a generar el informe.
	2.5.2.3.	Encriptar los archivos electrónicos de informes de resultados para almacenarlos y transferirlos.
	2.5.2.4.	Asegurarse de que los informes de resultados solamente sean enviados a los individuos que deban recibirlos. No enviar informes de resultados que sean más inclusivos de lo necesario. Puede ser más fácil volver a enviar el mismo informe completo a todos los usuarios del test, pero para mantener la confidencialidad de los candidatos, sólo se deben enviar los resultados relevantes a cada usuario del test.
	2.5.2.5.	Indicar a las instituciones que, para fines oficiales, sólo debe usarse el informe enviado directamente a la institución y no la copia que reciben quienes rinden el test (porque puede falsificarse). Asimismo, se debe recomendar a las instituciones que realicen verificaciones de rutina al informe institucional.

Finalmente, los investigadores han identificado una larga lista de buenas prácticas para la confección de informes individuales. Es emblemático en este sentido el trabajo de Goodman y Hambleton (2004), quienes estudiaron los informes de resultados y las guías interpretativas en Estados Unidos y Canadá. Luego de analizar estos documentos con un nivel de detalle sin precedentes, los autores identificaron buenas prácticas y realizaron recomendaciones generales para subsanar problemas de diseño y contenido. Entre sus recomendaciones se encuentran asegurarse de que los informes de resultados estén escritos de manera clara, concisa y visualmente atractiva, evitar la terminología estadística y tener en cuenta a los distintos usuarios al confeccionar los informes. Goodman y Hambleton también destacaron varias características problemáticas. Por ejemplo, notaron que muchos informes presentaban demasiada información sin referirse a puntos centrales, como por ejemplo indicadores de precisión o definiciones de términos claves. Ryan (2006) hizo eco de estos resultados al apuntar que muchos informes de resultados no incluyen información contextual sobre las puntuaciones y niveles de logro presentados, o bien no explican el desempeño de los estudiantes con un nivel de especificidad adecuado. Como respuesta a estos puntos, la

investigación ha examinado más detalladamente el modo de informar puntuaciones; así, la literatura hoy entrega una serie de recomendaciones para confeccionar informes generales de resultados (Goodman & Hambleton, 2004; National Education Goals Panel, 1998), informes para difusión en línea (Zenisky & Hambleton, 2012), informes de resultados de nivel grupal (Zenisky et al., 2009) e informes de resultados para padres de estudiantes de inglés (Zapata-River et al., 2014; Faulkner-Bond, Shin, Wang, & Zenisky, 2013). Es importante destacar que muchas de estas recomendaciones coinciden con los *Estándares* (AERA et al., 2014) y son tratadas en los modelos presentados por Zapata-Rivera (2011) y Hambleton y Zenisky (2013).

A pesar del aparentemente amplio consenso en el campo de la investigación sobre qué funciona y qué no en la generación de informes, existen aún muchos informes de resultados que presentan problemas (Zapata-Rivera, 2011). Uno de los problemas más críticos que afectan a los informes de resultados es el de la interpretabilidad. La investigación muestra que los informes de resultados son mal interpretados o difíciles de interpretar para variados públicos (Ward, Hattie, & Brown, 2003; Goodman & Hambleton, 2004; Ryan, 2006; Zenisky et al., 2009; Whittaker, Williams, & Wood, 2011; van der Kleij & Eggen, 2013). Otro problema común del que padecen estos informes es que no incluyen información sobre la precisión de las puntuaciones ni sobre el objetivo de la evaluación (Goodman & Hambleton, 2004). Este hallazgo es bastante sorprendente si se consideran los grandes esfuerzos realizados por asociaciones profesionales e investigadores para subrayar la importancia de explicitar estos aspectos. Los informes de resultados también suelen carecer de información sobre términos claves, a pesar de usar abundante terminología estadística (Goodman & Hambleton, 2004). Otros problemas observados en la literatura incluyen la escasa información contextual referida a descripciones de niveles de puntuación y logro (Ryan, 2006; Zenisky et al., 2009), la falta de información diagnóstica para relacionar los resultados de evaluación con la instrucción (NEGP, 1998) y la falta de informes sobre subgrupos (Zenisky et al., 2009), a pesar de la importancia capital de ciertas comparaciones grupales. Estos problemas reducen la utilidad de los informes de resultados; por ello, se les debe prestar atención para evitar que se desperdicien todos los recursos invertidos en el desarrollo de tests.

Como se ha esbozado, existen abundantes razones para afirmar que los informes de resultados siguen siendo problemáticos. Una posible explicación es que las agencias de evaluación no asignan suficientes recursos a la confección de informes, o bien que no toman en cuenta los asuntos relativos a los informes desde el comienzo del ciclo de evaluación (Zenisky & Hambleton, 2015). Luego, el problema podría estar relacionado con el proceso, por lo que los desarrolladores de tests debieran esforzarse más para mejorar sus prácticas de producción de informes. Otra posibilidad es que los desarrolladores de tests no estén conscientes de la gran cantidad de recursos disponibles para apoyar sus tareas de confección de informes. Por lo tanto, el problema podría estar relacionado con el conocimiento y debería ser posible resolverlo si la investigación se difundiera de modo más efectivo entre los desarrolladores de tests a cargo de los procesos de generación de informes. Una tercera posibilidad es que los esfuerzos dirigidos a la confección de informes no sean tan efectivos como desearían los desarrolladores de tests. Por ejemplo, los desarrolladores podrían no haber comprendido aún qué es lo que funciona realmente para cada público y para cada contexto específico de su programa de evaluación. Es probable que la investigación sobre los públicos principales no se haya realizado apropiadamente. Por otra parte, es posible que los públicos hayan ido cambiando

con el tiempo, por lo que sería necesario adaptarse a sus nuevas necesidades de información. Con todo, sigue existiendo en este campo una cierta distancia entre la cantidad de información y apoyo que entrega la literatura y los resultados de las prácticas actuales de los desarrolladores de tests en la confección de informes de resultados.

Evaluación de la calidad de los informes de resultados

Otra posible explicación de la distancia entre investigación y práctica con respecto a los informes de resultados es la falta de presión externa y la nula necesidad de rendir cuentas (accountability) en lo referido a las decisiones que influyen en la producción de informes. A los actores interesados (padres, profesores, comunidades escolares o legisladores, entre otros) les preocupa la calidad de los tests, pero es inusual encontrar discusiones públicas sobre los informes y sobre cómo éstos afectan la validez del uso de las puntuaciones. Incluso la documentación técnica parece restarle importancia a los informes en comparación a otros aspectos de la evaluación por medio de tests. Los manuales técnicos son muy descriptivos al referirse a temas de administración de tests o a las propiedades de constructos y puntuaciones, pero no enfatizan de igual modo la producción de informes: ¿por qué el informe presenta un cierto conjunto de información? ¿por qué son de un cierto tipo las decisiones de formato? Las decisiones referidas a los informes son críticas; sin embargo, este mensaje no ha sido necesariamente adoptado por el público.

La literatura psicométrica no entrega apoyo concreto para que los usuarios finales evalúen la calidad de los informes de resultados ni para que, por consiguiente, puedan exigir mejores informes. Pocos investigadores han desarrollado herramientas para la evaluación de informes. En su modelo de siete pasos, Hambleton y Zenisky (2013) presentan un formulario de evaluación para ser usado por los desarrolladores llegado el momento de hacer las modificaciones finales a los informes de resultados. El formulario HZ consiste en un grupo de preguntas dirigidas a reflexionar sobre los aspectos relevantes de los informes de resultados. El formulario HZ evalúa la calidad de los informes de resultados mediante 36 preguntas divididas en ocho dimensiones: (a) General: preguntas que se refieren al informe de manera holística; (b) Contenido - Introducción y descripción de informe: preguntas dirigidas a información relevante sobre la evaluación y/o programa que debe entregarse al comienzo del informe (por ejemplo, ¿el informe explica el propósito de la evaluación?); (c) Contenido - Puntuaciones y niveles de desempeño: preguntas relacionadas con la claridad de las escalas de puntuación incluidas en el informe; (d) Contenido - Otros indicadores de desempeño: preguntas sobre la claridad de las subescalas o de los análisis de los ítems, así como sobre los usos apropiados de esta información; (e) Contenido - Otros: preguntas sobre el apoyo entregado por las agencias a los usuarios finales para ayudarlos a interpretar y usar las puntuaciones; (f) Lenguaje: preguntas referidas a la terminología y el tono empleados en el informe; (g) Diseño: preguntas sobre la lógica según la cual se organizan los temas/preguntas, así como preguntas acerca de las decisiones de formato tomadas al confeccionar el informe y (h) Guías de interpretación y materiales adicionales: preguntas referidas al material empleado para apoyar la efectividad de los informes de resultados. El desarrollo del formulario HZ se basó en hallazgos de investigación y en buenas prácticas para la generación de informes: las preguntas y secciones incluidas en el formulario reflejan los aspectos más relevantes sugeridos por la literatura sobre informes de resultados. El Anexo A entrega más detalles sobre el formulario.

Gotch y Roberts (2014) hicieron esfuerzos adicionales para crear instrumentos que permitieran evaluar informes de resultados. Convencidos de que sistematizar la evaluación de informes de resultados podría ser beneficioso para la práctica psicométrica (Roberts & Gotch, 2016), estos autores desarrollaron y probaron su propio instrumento, basándose en gran medida en el formulario HZ y en su marco teórico. El objetivo de su trabajo fue desarrollar una herramienta para investigadores dirigida eventualmente a medir el impacto de la calidad de los informes de resultados sobre la validez de los usos de las puntuaciones obtenidas en tests (Gotch & Roberts, 2014; Roberts & Gotch, 2016). Con este objetivo en mente, tomaron el formulario HZ y eliminaron preguntas que consideraron no relevantes para los investigadores. Asimismo, transformaron las diferentes preguntas en afirmaciones, principalmente para permitir el uso de una escala de calificación. En total, los autores mantuvieron 31 criterios en 6 dominios.

Gotch y Roberts (2014; Roberts & Gotch, 2016) realizaron dos estudios distintos para desarrollar, evaluar y perfeccionar su instrumento. En primer lugar, los autores usaron su formulario para examinar 41 informes de resultados empleados en diferentes programas de evaluación estatales en los Estados Unidos. Su objetivo era probar el formulario y hacerle las modificaciones finales. Cada una de las 31 afirmaciones iba acompañada de una escala de evaluación de 3 puntos que correspondía a cada uno de los criterios siguientes: no cumplido (0), parcialmente cumplido (1) o totalmente cumplido (2). Los autores evaluaron personalmente los 41 informes y observaron que el formulario GR arrojaba resultados prometedores. Por una parte, las coincidencias exactas producidas por el formulario llegaron al 67% considerando todos los informes, mientras que se obtuvo acuerdo exacto + adyacente para el 94% de los criterios. Por otra parte, la escala de evaluación les permitió a los autores computar las puntuaciones promedio de cada informe. Los autores examinaron los informes que obtuvieron las puntuaciones más altas y más bajas, y descubrieron que los primeros eran efectivamente de mejor calidad que los segundos. También notaron que el formulario fue capaz de informar a los investigadores acerca de la variación específica entre informes de resultados en diferentes áreas de interés (Gotch & Roberts, 2014).

Dado que el formulario GR buscaba informar la investigación, una de sus consideraciones más importantes era la confiabilidad. Si el formulario es poco confiable, entonces no será adecuado para investigar el impacto de los informes de resultados sobre el uso válido de las puntuaciones de los tests. Por lo tanto, en su segundo estudio, los autores se centraron en investigar la confiabilidad del formulario GR. Basándose en su experiencia, los autores hicieron algunas modificaciones menores: emplearon una escala de evaluación de 4 puntos en lugar de 3 y entregaron descriptores específicos de cada punto de evaluación. Los autores les pidieron a 4 estudiantes de postgrado que evaluaran 3 informes de resultados con el formulario modificado. Luego, los autores llevaron a cabo un estudio de generalizabilidad de dos facetas en el cual los informes de resultados se cruzaron en su totalidad con los dominios y los observadores. Los dos hallazgos más importantes fueron que el formulario arrojó datos altamente confiables (G=0,78) y que, sin embargo, había otras fuentes sistemáticas y no identificadas de varianza que influían en la evaluación de los informes de resultados (Roberts & Gotch, 2016). Estos resultados son prometedores, puesto que indican que el formulario GR tiene el potencial para sistematizar la evaluación de la calidad de los informes de resultados en el campo de la investigación científica.

Propósito

La literatura psicométrica incluye sólo dos instrumentos para evaluar la calidad de los informes de resultados: el formulario HZ, dirigido a facilitar la labor de los desarrolladores de tests, y el formulario GR, que busca apoyar a los investigadores. El formulario HZ fue desarrollado para sistematizar y mejorar el proceso de creación de los informes de resultados (Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015). Por su parte, el formulario GR fue desarrollado para permitir el análisis sistemático de los informes de resultados y su impacto sobre las consideraciones de validez (Gotch & Roberts, 2014; Roberts & Gotch, 2016). Una de las consecuencias indirectas del formulario GR es empoderar a la comunidad de investigación, permitiendo que los investigadores lo usen para estimular y exigir la generación de mejores informes de resultados. Creemos que esta es una contribución valiosa a este campo; sin embargo, podría capitalizarse mejor si fuéramos capaces de empoderar a los usuarios finales (por ejemplo, padres, profesores y comunidades escolares) para que evalúen la calidad de los informes y para que, como producto de ello, puedan realizar peticiones similares a los desarrolladores de tests. Coincidimos completamente con la idea que los informes de resultados son esenciales para usar e interpretar las puntuaciones de los tests de manera válida (Hattie, 2009; Roberts & Gotch, 2016), pero consideramos que dicha postura prácticamente no existe en los debates del público general referidos a la evaluación de los tests y su impacto. Si bien existe abundante literatura sobre el tema de los informes de resultados, la investigación no ha producido herramientas tangibles que eduquen a los actores interesados y los empoderen para que sean capaces de exigir informes de mejor calidad. El propósito de este estudio es dar un primer paso en esa dirección.

El formulario de evaluación creado por Hambleton y Zenisky (2013) es un buen punto de partida para desarrollar una herramienta de este tipo. El formulario HZ se basa en investigación científica, es exhaustivo y se centra en aspectos que ciertamente son relevantes para muchos actores interesados. Así, coincidimos con Gotch y Roberts (2014) en cuanto a que el formulario HZ sirve como base para el desarrollo de otros formularios de evaluación, incluyendo uno que permita educar y empoderar a los usuarios finales. El objetivo específico de este artículo es determinar cuán apropiado y útil es el formulario HZ (en su forma original) para evaluar informes de resultados desde la perspectiva de los usuarios finales. Con este propósito en mente, examinamos la claridad, usabilidad y significación del formulario HZ para evaluar un informe de resultados de nivel individual típico. Nuestros hallazgos, junto con los de otros estudios similares, informarán dos estudios consecutivos orientados a desarrollar un nuevo formulario que permita que los usuarios finales evalúen informes de resultados de manera independiente y en una amplia gama de contextos.

Método

Para examinar la claridad, usabilidad y significación del formulario HZ como herramienta para evaluar informes de resultados, se llevó a cabo un estudio basado en grupos focales a pequeña escala empleando una muestra intencional.

Participantes

Participaron seis estudiantes de postgrado (tres hombres y tres mujeres) quienes se presentaron voluntariamente a un grupo focal. Todos los participantes eran alumnos de una Escuela de Educación en una universidad estadounidense y, por ello, estaban relativamente al tanto de los conceptos esenciales de la evaluación educacional. Más aún, una breve encuesta demográfica reveló que todos los participantes recordaban haber recibido al menos un informe de resultados que describía su propio desempeño en un test estandarizado y que todos habían visto en promedio 3,5 informes de resultados únicos en los últimos cinco años.

Materiales

Los participantes evaluaron un informe de resultados tipo usando una versión en papel del formulario HZ (ver Anexo A). El instrumento se formateó como una tabla de dos columnas, en la cual las preguntas se presentaban en la primera columna y se dejaban espacios en blanco para responder en la segunda columna.

El informe de resultados (ver Anexo C), el cual mostraba el desempeño de un estudiante ficticio en una evaluación de la asignatura de Inglés en una escuela secundaria, fue seleccionado de un grupo de muestras de informes de resultados recolectado en los sitios web de los departamentos de educación de más de 25 estados. Se siguió un meticuloso proceso de selección para asegurar que las idiosincrasias del informe que se usara no tuvieran un papel indebido en el formulario HZ. El informe fue finalmente seleccionado porque representa un documento de "calidad media" de acuerdo al juicio holístico de los autores e incluye contenidos y elementos de diseño claves cada vez más comunes en los informes de resultados modernos, como indicadores de desempeño en sub-áreas, un gráfico de desempeño global y un texto interpretativo de la puntuación global y las puntuaciones de cada sub-área. (Debe tenerse en cuenta que la información de identificación se ha difuminado u omitido en la figura presentada en el Anexo C, pero la versión del informe de resultados usada en los grupos focales no fue alterada).

Procedimiento

Recolección de datos. Los participantes fueron asignados a dos grupos de tres; luego, se le pidió a cada grupo que asistiera a un grupo focal de una hora de duración. En cada reunión, los participantes llevaron a cabo un ejercicio de 20 minutos de duración en el cual debían usar el formulario HZ para evaluar un informe de resultados de muestra. Luego de usar el formulario, los participantes conversaron durante 30 minutos sobre su experiencia al emplearlo. La conversación fue semiestructurada y estuvo dirigida a obtener retroalimentación sobre temas centrales de este estudio. Las conversaciones del grupo focal fueron grabadas (con el consentimiento de los participantes) y transcritas. En el Anexo B se entregan más detalles sobre el formato y las preguntas del grupo focal.

Análisis de datos. Primero, se llevó a cabo un análisis cualitativo de las transcripciones de los grupos focales. El objetivo fue identificar los temas que a los participantes les parecieron más relevantes y resumir sus visiones sobre cada uno de ellos. Este procedimiento no incluyó temas

predefinidos. A continuación, se analizaron las respuestas de los participantes a las preguntas incluidas en el formulario HZ. Se analizaron estos datos para comprender hasta qué punto el formulario HZ era claro, fácil de usar y significativo para los participantes al evaluar el informe de resultados que se les había entregado. Específicamente, sobre la base de las respuestas obtenidas, se calcularon las frecuencias de los siguientes tipos de preguntas: (a) las que les parecieron confusas a los participantes, (b) las que provocaron respuestas consistentes en los participantes, (c) las que, en lugar de ser abiertas, podrían haber empleado una escala de evaluación, (d) las que los participantes no lograron responder porque necesitaban actividades de seguimiento y (e) las que los participantes consideraron redundantes. Por ejemplo, para informar nuestra evaluación de la claridad del formulario HZ, una de las frecuencias que calculamos indicó cuántas preguntas fueron confusas para los participantes, considerando respuestas como "No estoy seguro" y "No entiendo."

Resultados

Análisis cualitativo de las transcripciones

La primera parte de los análisis consistió en inspeccionar las transcripciones de los grupos focales. Como se explica más adelante, emergieron varios temas.

El primer tema que surgió de las discusiones de los grupos focales fue la poca certeza de los participantes con respecto a los actores interesados (1): ¿Para quién estamos evaluando esto? ¿Y quién va a usar este formulario? Sin esta información, fue imposible responder algunas de las preguntas del formulario HZ, como la pregunta I.B, que dice "¿Refleja el informe de resultados los intereses y necesidades de información de los actores interesados claves?" Sin saber quiénes son estos actores interesados, las preguntas les parecieron ambiguas a los participantes, quienes se sintieron incapaces de contextualizar sus respuestas.

> *"Cuando vi por primera vez la pregunta sobre los actores interesados, pensé 'este es un informe de resultados para familias', así que pensé que los actores interesados serían solamente los padres o los estudiantes que rindieran un test, pero usualmente los actores interesados también incluyen legisladores o profesores, así que me confundí un poco con esa pregunta."*

El objetivo del formulario HZ era entregar información a desarrolladores de tests, quienes sabrían de antemano quiénes eran los actores interesados. A primera vista, las preguntas referidas a los actores interesados podrían parecer redundantes o poco pertinentes dado que el objetivo del formulario que queremos desarrollar es informar a los propios actores interesados. Sin embargo, en esta etapa buscamos evitar cualquier ajuste basado en nuestro propio juicio para tener una idea de cómo reaccionaban los participantes a cada pregunta.

Un segundo tema que emergió se refiere a la imagen del informe de resultados ideal (2) que crearon los participantes a partir de lo que sugiere el formulario HZ. Sobre la base de las preguntas, los participantes imaginaron que un informe abarrotado y lleno de información sería el ideal. Es interesante observar que Goodman y Hambleton (2004) hacen una advertencia específica con respecto a esta situación; en este sentido, una sugerencia común es que los informes de resultados sean concisos y ordenados. Así, algunas partes del formulario causaron una impresión errada de lo que debiera ser un buen informe de resultados, aunque la mayoría de las sugerencias coincide con lo

que para los investigadores constituye un buen informe de resultados.

Un tercer tema surgido de los grupos focales fue que a los participantes les pareció poco pertinente el contenido de las preguntas (3). En primer lugar, los participantes mencionaron que muchas preguntas eran redundantes y que consideraban que debían dar la misma respuesta varias veces. Otro grupo de comentarios tuvo que ver con la sensación de los participantes de que algunas preguntas no eran verdaderamente importantes (aunque sí concedieron que, antes de dar una opinión tajante, era necesario saber para qué se estaba evaluando el informe de resultados). Asimismo, los participantes consideraron que las preguntas del formulario HZ dejaban fuera algunos puntos importantes, como por ejemplo los elementos gráficos de los informes. El informe evaluado por los participantes (ver Anexo D) tenía un gráfico y los participantes hicieron numerosos comentarios sobre él. En particular, dijeron que el gráfico les pareció confuso (por ejemplo, la escala no estaba clara) y que no podían expresar esto mediante el formulario HZ.

Con respecto a este tema, los participantes también mencionaron que el lenguaje de las preguntas les pareció anticuado. Además, hicieron notar que algunas preguntas eran muy fáciles mientras que otras requerían algún tipo de conocimiento experto. En otras palabras, sintieron que las preguntas eran muy variadas en cuanto a dificultad. Asimismo, algunas preguntas no podían responderse sin información adicional (por ejemplo, "¿Hay alguna guía interpretativa preparada? De ser así, ¿es informativa y está escrita de forma clara?"). Los participantes se mostraron un poco confundidos sobre si debían o no "deducir" parte de esta información faltante basándose en el informe.

> *"Aparte de la pregunta sobre los actores interesados, creo que hay una pregunta sobre 'informes o materiales disponibles en diferentes medios, si coinciden o no con los materiales relacionados que se han publicado', esas preguntas, no sé cómo se podrían responder sólo sobre la base de esto, así que se necesita más información. Y después, al incluir el promedio de la escuela, del distrito y del estado, no sé si querían inducirnos a decir cuáles eran los actores interesados."*

Un cuarto tema que emergió de la conversación fue que ciertas preguntas catalizaron reflexiones significativas sobre el informe de resultados (4), lo cual se percibió como algo positivo. Este es un hallazgo interesante, dado que nuestra intención es desarrollar un formulario que invite a los actores interesados a pensar críticamente sobre los informes.

> *"Creo que hubo algunas preguntas que nos guiaron en direcciones útiles. En realidad no había procesado la escala del evaluador, así que cuando me preguntaron sobre eso, dije '¡es cierto... eso no tiene sentido!'"*

El quinto punto derivado de la conversación tuvo que ver con el formato de las preguntas del formulario HZ: los participantes consideraron que el formato de las opciones de respuesta no siempre era apropiado (5). El formulario HZ original consiste en un conjunto de preguntas sin ninguna escala asociada. En este estudio, añadimos un espacio en blanco junto a cada pregunta para que así los participantes pudieran responder en el formato de su preferencia. Algunos participantes se quejaron porque el recuadro para escribir sus respuestas era demasiado pequeño; sin embargo, este no es un problema del formulario HZ sino del impreso entregado a los participantes, el que fue producto de una decisión nuestra. Aun así, sin importar el tamaño concreto de los cuadros de

respuesta, algunos participantes manifestaron que las expectativas de respuesta no estaban claras, puesto que algunas preguntas podían responderse con una sola palabra (por ejemplo "sí" o "no") mientras que otras requerían explicaciones más detalladas, a pesar de que todas las preguntas tenían el mismo formato.

> *"Sentí que no tenía realmente claro cuánto querían saber. Entonces no pude terminar la primera pregunta porque el cuadro era demasiado pequeño, pero en otras preguntas me dio la impresión de que solamente había que responder sí o no. ¿Quieren una respuesta súper larga o no?"*

Los participantes sugirieron que se usara una escala Likert en algunas preguntas, lo que simplificaría el uso del formulario y potencialmente lo haría más útil. Además, los participantes mencionaron que el carácter subjetivo de las preguntas sugería opciones de respuesta del tipo "Creo que sí" en vez de otras más objetivas como "sí/no".

El sexto tema que surgió fue el de los problemas estructurales del formulario (6). Se presentaron algunas sugerencias con respecto a este punto. Una primera sugerencia de los participantes fue cambiar el orden de las preguntas, para que así el formulario comenzara con las preguntas simples (por ejemplo, las referidas a temas de diseño) y terminara con las globales o más generales/complejas. Les pareció más lógico dejar las preguntas sumativas para el final, después de haber reflexionado sobre todos los problemas específicos del formulario. Una segunda sugerencia fue eliminar los encabezados entre grupos de preguntas, ya que ocupaban espacio y no se les prestaba atención. Una tercera sugerencia fue incluir una tabla de contenidos al comienzo del formulario para que los participantes tuvieran una idea de los temas y las preguntas que se incluyen. De este modo, los usuarios podrían planificar mejor sus respuestas.

> *"… estaba pensando que ojalá hubiéramos tenido una tabla de contenidos, o sea algo como 'estas son las preguntas que vamos a ver', porque así en vez de escribir lo mismo varias veces habría guardado algunas respuestas para la [sección] más relevante."*

Un séptimo tema que surgió fue la claridad de las preguntas, ya que los participantes consideraron que las preguntas eran confusas (7). En su opinión, la confusión tenía que ver con el vocabulario usado: sugirieron que usar palabras más simples resolvería el problema.

El último problema mencionado en los grupos focales fue que el formulario HZ, en su versión actual, no es totalmente útil para entregar una evaluación global de un informe de resultados (8). Por una parte, sin opciones de respuesta dentro de una escala, es imposible entregar una puntuación sumativa que permita hacer un juicio final. Incluso con opciones de respuesta dentro de una escala de evaluación, es difícil combinar las preguntas de diferentes secciones. Por otra parte, emplear este formulario requiere algún tipo de especialización (por ejemplo, conocimientos de psicometría): es poco probable que una persona por sí sola pueda usar el formulario para juzgar la calidad de un informe de resultados. Los participantes no minimizaron la relevancia del formulario para fines de evaluación, pero consideraron que había sido construido con un informe particular en mente, lo que no necesariamente ayudaba a los usuarios finales a evaluar informes.

> *"Me pareció que la pauta había sido creada con un informe de resultados específico en mente, y estoy seguro de que si hubiéramos estado viendo ese informe de resultados algunas preguntas habrían tenido mucho más sentido."*

Análisis de las respuestas de los participantes

Además de las transcripciones de los grupos focales, analizamos las respuestas de los participantes al formulario HZ, haciendo énfasis en la claridad, utilidad y significación de éste.

Sobre la claridad del formulario HZ. ***¿Cuán claras fueron las preguntas y secciones del formulario HZ?*** Para hallar una respuesta, examinamos el número de preguntas que a los participantes les parecieron confusas, de acuerdo a dos criterios: las que tenían respuestas del tipo "No entiendo la pregunta" y las que tenían respuestas que indicaban que los participantes efectivamente estaban confundidos. Añadimos estas dos categorías para calcular la frecuencia de las respuestas que resultaron confusas para los participantes, considerando cada pregunta individualmente. Definimos tres niveles de confusión: no confusas – preguntas que no confundieron a los participantes; poco confusas – preguntas para las que encontramos de 1 a 3 respuestas (de un total de 6) que indicaban confusión y confusas – preguntas para las cuales hubo 4 o más respuestas que indicaban confusión. En general, las preguntas fueron no confusas (78%) o poco confusas (19%), como muestra la Figura 1.

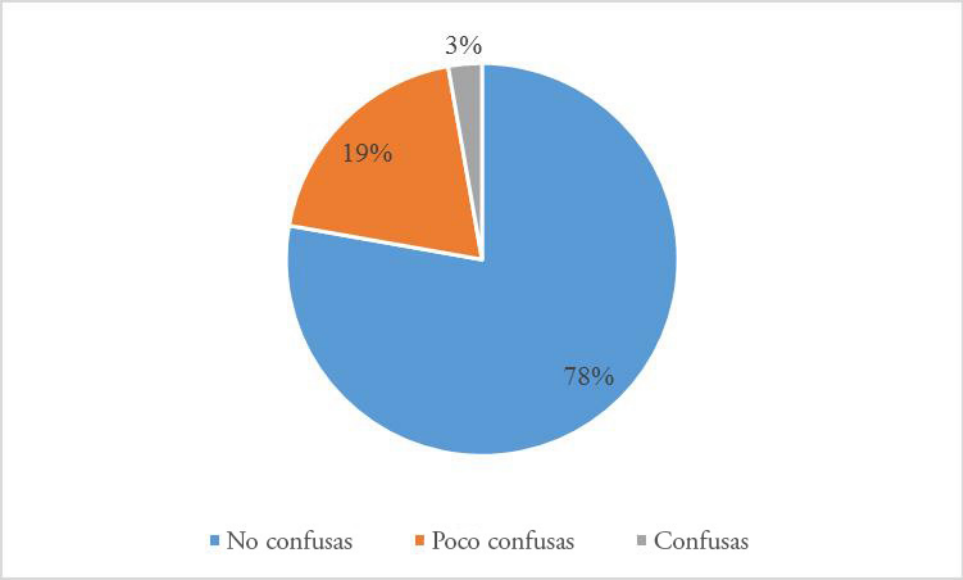

Figura 1. Preguntas que resultaron confusas para los participantes

¿En qué grado el formulario HZ generó respuestas consistentes? Para hallar una respuesta, se examinó el grado en el cual cada pregunta llevó a los participantes a sacar conclusiones similares, considerándose que respuestas como "sí", "me parece que sí" y "en gran medida" ilustraban la misma conclusión. Se aplicó la misma regla a las expresiones que señalaran conclusiones negativas (por ejemplo, "no", "no creo" y "para nada"). Los conjuntos de respuestas a cada pregunta se evaluaron como consistentes si todos los participantes llegaban a la misma conclusión, parcialmente consistentes si cuatro de seis participantes llegaban a la misma conclusión e inconsistentes si menos de cuatro participantes llegaban a la misma conclusión. Como se observa en la Figura 2, se encontró que 6 preguntas generaron conclusiones consistentes, 16 generaron conclusiones parcialmente consistentes y 14 generaron conclusiones inconsistentes.

Cabe mencionar que la consistencia de las respuestas está influida por las características de las preguntas y de los participantes. Otro grupo de evaluadores podría entregar respuestas mucho más (o mucho menos) consistentes. Más aún, si bien la consistencia puede ser práctica cuando se realiza una afirmación sumativa de la calidad de un informe de resultados, un cierto nivel de inconsistencia indica que el formulario HZ es capaz de generar distintas perspectivas. La revisión de las respuestas calificadas como parcialmente consistentes o inconsistentes mostró que los siguientes tipos de pregunta tenían mayores probabilidades de recibir respuestas inconsistentes: las preguntas muy abiertas ("¿Cuáles son tus *impresiones generales* sobre el informe?"), las que requieren conocer la opinión de otros individuos ("¿El informe de resultados refleja los intereses informativos de los *principales actores interesados*?"), las que se refieren a múltiples aspectos (¿El informe indica *números telefónicos, sitios web o direcciones físicas* donde puedan hacerse consultas?) y las que se perciben como confusas ("¿Hay información que describa la *unidad de análisis* que se reporta?").

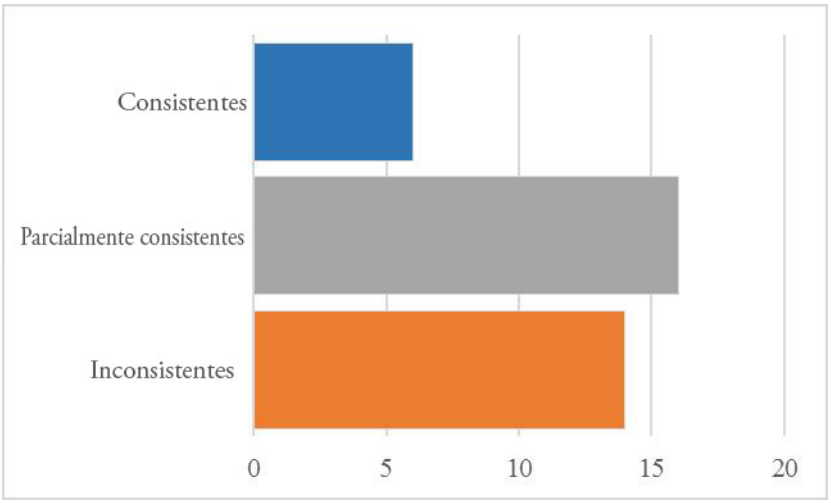

Figura 2. Niveles de consistencia de 36 preguntas

Sobre la usabilidad del formulario HZ. *¿Cuán adecuado es el formato de las preguntas del formulario HZ?* Uno de los temas que surgió en los grupos focales fue la pertinencia del formato de las preguntas. Este punto también fue mencionado por Gotch y Roberts (2014), quienes terminaron por incluir una escala fija para acompañar las afirmaciones de su formulario. Para comprender si este asunto afectaba sólo a algunas preguntas, a la mayoría de ellas o a todas, se se calculó la frecuencia de aquellas que podrían tener una respuesta sí/no (ver Figura 3) o contestarse con una escala de evaluación (ver Figura 4). Según se observó, pocas preguntas recibieron respuestas del tipo sí/no (24%), pero la mayoría (58%) se respondió como si se hubiera usado una escala de evaluación, lo que sugiere que la mayor parte de las preguntas se beneficiaría con la inclusión de una escala fija.

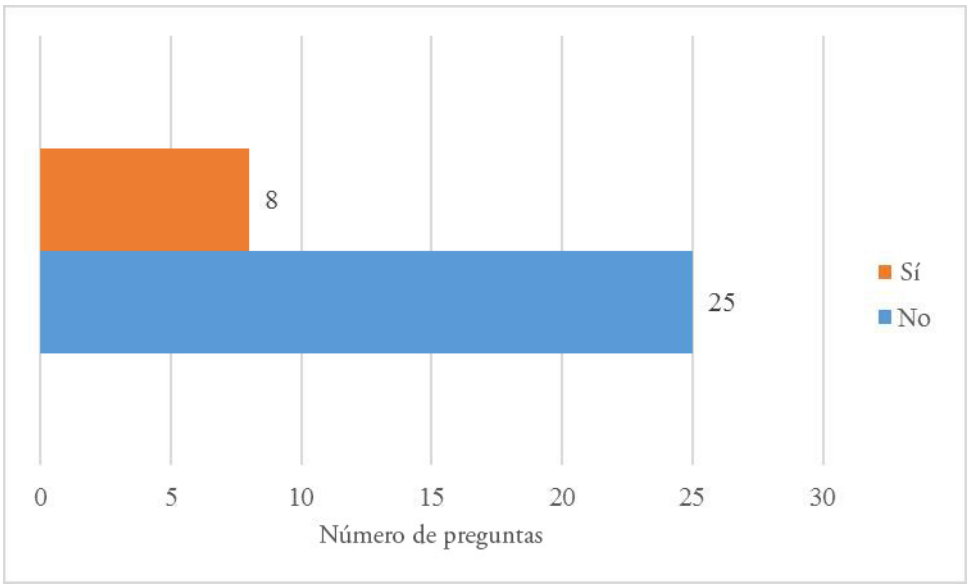

Figura 3. Preguntas que recibieron respuestas sí/no

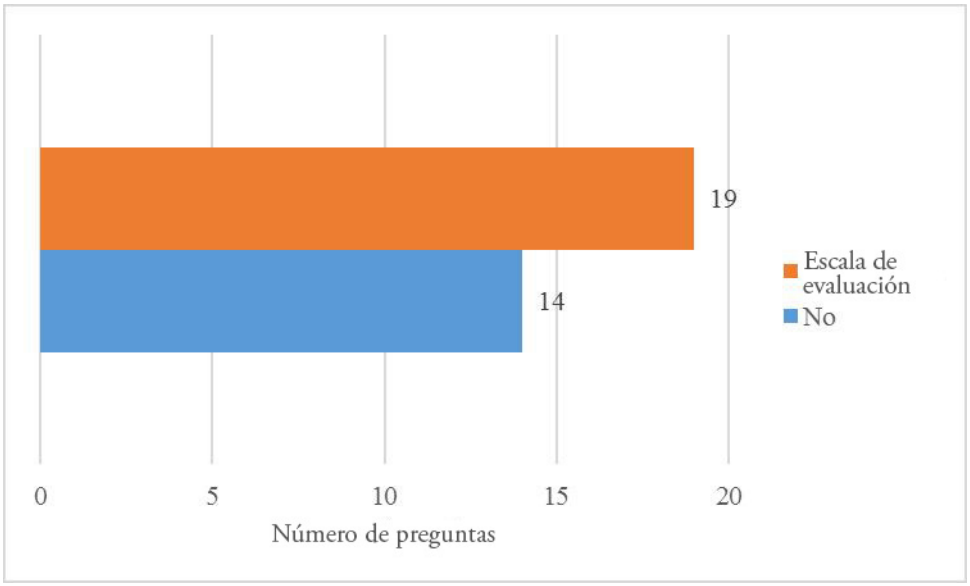

Figura 4. Preguntas respondidas como si se emplearan escalas de evaluación

¿Hasta qué punto las preguntas del formulario HZ requieren conocimientos profundos de estadística o diseño por parte de los evaluadores? Los participantes tenían la opción de omitir sus respuestas a algunas preguntas si sentían que necesitaban algún tipo de conocimiento experto (por ejemplo, sobre estadística o diseño). Así, logramos calcular la frecuencia de las preguntas que requerían conocimientos técnicos. Como se observa en la Figura 5, los participantes consideraron que sólo el 14% de las preguntas requerían algún grado de conocimiento experto.

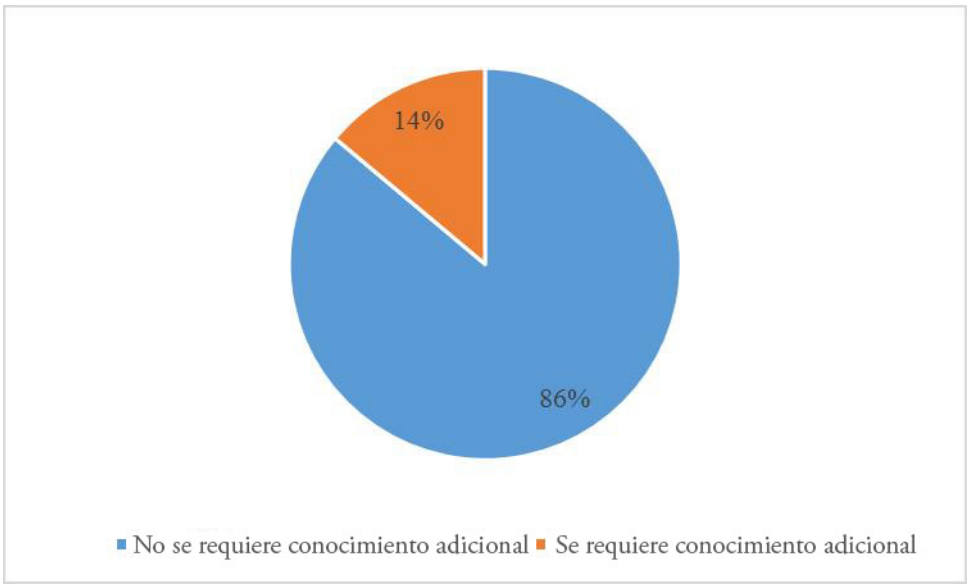

Figura 5. Nivel de conocimiento requerido por las preguntas

Sobre la significación del formulario HZ. ***¿Hasta qué punto las preguntas del formulario HZ necesitaban actividades de seguimiento?*** Los participantes también tenían la opción de no contestar algunas preguntas si sentían que dar una respuesta adecuada requería algún tipo de investigación o actividad de seguimiento. Un análisis de frecuencias simple aplicado a las respuestas indica que el 64% de las preguntas no requería ningún tipo de actividad de seguimiento, mientras que el 25% de requería actividades de este tipo en un grado medio (de 1 a 3 participantes lo consideraron así) y el 11% necesitaba un alto nivel de actividades de seguimiento (de 4 a 6 participantes opinaron de este modo).

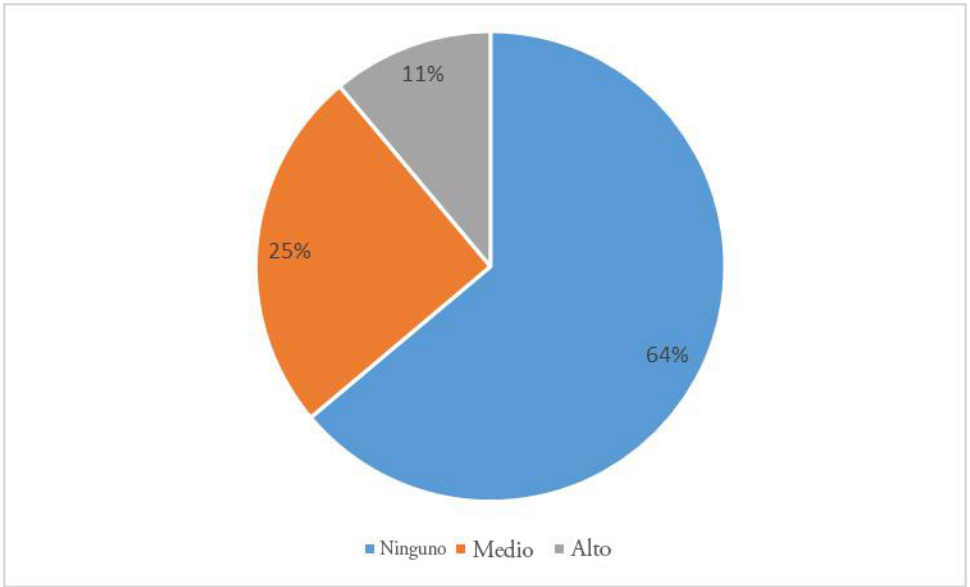

Figura 6. Nivel de actividades de seguimiento

¿Cuán redundantes eran las preguntas del formulario HZ? Si bien los participantes mencionaron que el formulario contenía algunas preguntas redundantes, sus respuestas indicaron que sólo dos (6%) podían clasificarse de ese modo. En el caso de estas dos preguntas, algunos participantes entregaron respuestas que partían por "véase respuesta anterior" o "nuevamente…", aunque las respuestas de este tipo estuvieron prácticamente ausentes.

Discusión

Los informes de resultados son esenciales para el uso válido de las puntuaciones de los tests, dado que son el único medio que conecta a los desarrolladores de tests con los actores interesados. Si los actores interesados no son capaces de entender o usar la información contenida en un informe de resultados, todos los esfuerzos y recursos empleados en el desarrollo de tests y en la recolección de datos serán en vano. Por ello, durante los últimos 10 años los investigadores han dirigido su atención a la confección de informes y han generado múltiples recursos que buscan mejorar los procesos de entrega de información. En consecuencia, y desde entonces, las prácticas de entrega de información han avanzado, especialmente con respecto al proceso de confección de informes. La investigación ha tenido un impacto positivo y efectivo en la práctica psicométrica.

A pesar de todos los avances logrados, los usuarios finales parecen no estar conscientes de cuán importantes son los informes de resultados ni de la relevancia de exigir informes más claros, útiles y significativos. El debate público referido a los tests muy rara vez se centra en temas de entrega de información, ya que (erróneamente) se da por descontado el conocimiento sobre datos psicométricos y de desempeño. Sin embargo, la interpretación de datos psicométricos es un proceso complejo cuya importancia no debe ser minimizada. Si bien reconocemos que los desarrolladores de tests tienen en cuenta a los actores interesados e implementan múltiples ciclos de retroalimentación para mejorar los informes finales, creemos que esto no basta. Hay muchos contextos en los cuales los informes se siguen dejando para el final del desarrollo de tests, con la consiguiente falta de cuidado en su producción. Adicionalmente, los informes pueden ser poco claros o inútiles para muchos actores interesados, cuyos conocimientos sobre evaluación pueden variar de forma sustantiva. Los actores interesados merecen recibir informes de resultados de buena calidad y deben estar conscientes tanto de su importancia como de las cualidades que caracterizan a los buenos informes. En ese sentido, la investigación se ha quedado rezagada.

Una herramienta accesible que les permita a los usuarios finales juzgar la calidad de los informes de resultados es muy necesaria para que ellos (por ejemplo, padres, profesores o analistas de políticas públicas) se den una idea de la buena o mala calidad de los informes. Este estudio representa un primer paso en esta dirección ya que evalúa la claridad, utilidad y significación del formulario de Hambleton y Zenisky (2013). El formulario HZ no fue pensado para usuarios finales, pero constituye una base fuerte para iniciar este trabajo. Así, obtuvimos retroalimentación de seis alumnos de postgrado para comprender hasta qué punto el formulario HZ es útil para evaluar un informe de resultados típico y qué modificaciones se requerían. Sobre la base de sus respuestas y conversaciones, es posible definir una serie de pasos siguientes dirigidos a producir un formulario que permita que los usuarios finales juzguen la calidad de los informes de resultados.

Nuestra primera conclusión es que el contenido del formulario HZ no es apropiado para usuarios finales. Las transcripciones de los grupos focales muestran que muchas preguntas les parecieron redundantes a los participantes, algo similar a lo que apuntaran Gotch y Roberts (2014). El contenido de algunas preguntas también era poco apropiado debido a problemas de lenguaje (por ejemplo, términos anticuados o excesivamente complicados). Los participantes también mencionaron que las preguntas no tenían en cuenta algunos puntos importantes del informe, mientras que otras fueron consideradas irrelevantes. Estos hallazgos no son necesariamente observables en las respuestas escritas de los participantes, ya que solamente el 6% de las preguntas fue clasificado como irrelevante y el 28% como un poco confuso. Sin embargo, las respuestas indican que sólo el 17% de las preguntas generaron respuestas consistentes, lo que significa que los participantes tendieron a interpretar las preguntas de modo bastante variable. Toda esta información indica que las preguntas deben mejorarse en cuanto a lenguaje, redundancia y claridad, pero también que los usuarios finales necesitan más orientación.

El asunto de la orientación se relaciona con otra queja de los participantes: que las preguntas deberían tener escalas (o que no todas deben ser abiertas). Globalmente, el 58% de las preguntas podría haberse escrito fácilmente como preguntas con escala, lo que habría hecho que el ejercicio de evaluación fuera más práctico. Los participantes sintieron que las escalas de evaluación les habrían permitido sacar mejores conclusiones sobre la calidad del informe, algo que también mencionan Gotch y Roberts (2014). Sin embargo, esta recomendación no puede ser implementada directamente en un instrumento para usuarios finales, ya que éstos no desean comparar informes de resultados entre programas ni hacer evaluaciones consistentes; más bien, los usuarios finales quieren entender qué aspectos de los informes deben tener en cuenta para así entender, en base a ellos, qué es lo que define a un buen informe en un contexto dado. Creemos que usar escalas de evaluación sería beneficioso para las preguntas que requieren conocimiento de tipo declarativo, pero que las preguntas más holísticas debieran seguir siendo abiertas. El objetivo de una herramienta de evaluación para usuarios finales sería no sólo educar a los actores interesados mediante la inclusión de preguntas sobre aspectos que son relevantes en todo informe de resultados, sino también empoderarlos permitiéndoles desarrollar un juicio propio e independiente sobre los informes de resultados. Consideramos que una combinación de preguntas cerradas y abiertas serviría para alcanzar ambos propósitos.

Sin embargo, el orden de las preguntas es relevante. Una de las sugerencias más interesantes que recibimos de los grupos focales tuvo que ver con la dificultad de las preguntas: para los participantes, debieran ordenarse de forma progresiva, partiendo por las más fáciles y terminando por las más difíciles. Esto tiene sentido, ya que algunas preguntas del formulario HZ estimulan reflexiones sobre los informes que podrían capitalizarse mejor en un juicio final realizado al terminar la actividad. Sin embargo, algunos participantes mencionaron que la imagen del informe de resultados ideal que se desprendía de este formulario era ambigua y a veces no pertinente. Sin importar cuales sean, todas las preguntas o afirmaciones incluidas en una herramienta final deben apuntar claramente a los aspectos relevantes de un informe de resultados. Además, para aumentar la significación de este formulario, es esencial simplificar ciertas preguntas que actualmente parecen necesitar conocimiento experto de algún tipo (14%) y eliminar aquellas que requieren actividades de seguimiento (36%), excepto cuando se trate de actividades relacionadas directamente con los usuarios finales.

El paso que viene inmediatamente a continuación de este estudio es el desarrollo de un prototipo de un instrumento adaptado. Inicialmente, proponemos las siguientes modificaciones al formulario HZ: incluir una tabla de contenidos, modificar los encabezados actuales de cada sección, cambiar el orden de las preguntas para que las más holísticas queden al final, eliminar preguntas que sean redundantes o irrelevantes para los usuarios finales, reescribir las preguntas como afirmaciones claras, usar escalas de evaluación cuando sea pertinente, entregar más orientación para contestar cada pregunta (incluyendo ejemplos o usando escalas de evaluación claramente definidas) y cerciorarse de que las preguntas apunten a la imagen correcta de lo que debiera ser un buen informe de resultados. Sin embargo, visualizamos múltiples formatos posibles para esta nueva herramienta, los que necesitarían retroalimentación adicional. De hecho, sería necesario obtener retroalimentación de un grupo más grande y variado de participantes, los que debieran evaluar un conjunto mayor de informes. Reconocemos que las principales limitaciones del presente estudio tienen que ver con su escasa amplitud: empleamos un pequeño grupo de participantes que podrían no ser representativos de la idea que tenemos de los usuarios finales y además trabajaron con un único informe de resultados. A pesar de ello, dicha limitación no es crítica en esta etapa, ya que este es el primer paso de un programa de investigación más amplio.

El artículo original fue recibido el 27 de diciembre de 2016
El artículo fue aceptado el 30 de octubre de 2017

Referencias

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

Faulkner-Bond, M., Shin, M., Wang, X., Zenisky, A., & Moyer, E. (2013, April). Score reports for English proficiency assessments: Current practices and future directions. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17(2), 145-220. doi: 10.1207/s15324818ame1702_3

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), American Psychological Association Handbook of Testing and Assessment in Psychology (pp. 479-494). doi: 10.1037/14049-023

International Test Commission (ITC). (2000). International guidelines for test use. Retrieved from http://www.intestcom.org/guidelines/index.php: International Test Commission (ITC).

International Test Commission (ITC). (2012). ITC Guidelines for quality control in scoring, test analysis, and reporting of test scores. Retrieved from http://www.intestcom.org/guidelines/index.php

Jaeger, R. M. (2003). NAEP validity studies: Reporting the results of the National Assessment of Educational Progress (Working Paper 2003-11). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Knupp, T., & Ansley, T. (2008, March). Online, state-specific assessment score reports and interpretive guides. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

National Education Goals Panel (NEGP). (1998). Talking about tests: An idea book for state leaders. Washington, DC: U.S. Government Printing Office.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. W. Downing & T. M. Haladyna (Eds.), Handbook of Test Development (pp. 677-710). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. Studies in Educational Evaluation, 39, 144-152. doi: 10.1016/j.stueduc.2013.04.002.

Ward, L., Hattie, J. A. C., & Brown, G.T. (2003). The evaluation of asTTle in schools: The power of professional development. AsTTle technical report 35, University of Auckland/New Zealand Ministry of Education.

Whittaker, T. A., Williams, N. J., & Wood, B. D. (2011). Do examinees understand score reports for alternate methods of scoring computer based tests? Educational Assessment, 16(2), 69-89.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), Test Score Reporting: Perspectives From the ETS Score Reporting Conference. (ETS Research Report No. RR-11-45). Princeton, NJ.

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. Assessment In Education: Principles, Policy & Practice, 21(4), 442-463.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. Educational Measurement: Issues and Practice, 31(2), 21-26.

Zenisky, A. L., & Hambleton, R. K. (2015). Good practices in score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), Handbook of test development (pp. 585-602). New York, NY: Routledge.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. Applied Measurement in Education, 22(4), 359–375. doi:10.1080/08957340903221667

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. Educational Assessment, 19(2), 116-138. doi: 10.1080/10627197.2014.903653

Apéndice

A. Formulario de evaluación de Hambleton y Zenisky (Hambleton & Zenisky, 2013)

Elemento del informe		Preguntas de evaluación del informe de resultados
I. Globales	A.	¿Cuáles son tus impresiones generales sobre el informe?
	B.	¿El informe de resultados refleja los intereses y necesidades de información de los actores interesados claves?
II. Contenido - Introducción y descripción del informe	A.	¿El informe tiene un título que indique claramente qué es?
	B.	¿Se entregan detalles sobre el contenido del test/los tests sobre el cual/los cuales se está informando?
	C.	¿Hay información que describa la unidad de análisis que se reporta?
	D.	¿Se describe el propósito / los propósitos del test?
	E.	Si es que se incluye, ¿se establece un tono positivo para el informe en la presentación de la agencia patrocinante (por ejemplo, gobernador, presidente, psicólogo, etc.)?
III. Contenido - Puntuaciones y niveles de desempeño	A.	¿Se explicita el rango de la escala de puntuación?
	B.	¿Se describen de forma suficiente para el público objetivo las categorías de desempeño o los estados psicológicos usados (por ejemplo, no logrado, básico, competente, avanzado, logrado)?
	C.	¿Se entrega información sobre cómo usar y no usar todas las puntuaciones y clasificaciones numéricas?
	D.	¿Se entregan ejemplos concretos del uso correcto de las puntuaciones informadas?
	E.	¿Se toca el tema de la imprecisión de las puntuaciones para cada puntuación que se informa? Esto puede hacerse con descripciones, gráficos o cifras.
	F.	¿Se evitan las "probabilidades" o las "probabilidades condicionales"? Si se usan, ¿se entrega una explicación clara?
	G.	¿Se evitan las notas a pie de página? Si se usan, ¿están escritas claramente para el lector?
	H.	¿Hay suficiente información para el lector, sin llegar a abrumarlo?
IV. Contenido - Otros indicadores de desempeño	A.	¿Se asocian los resultados del test con posibles actividades de seguimiento? Por ejemplo, en el caso de tests en el ámbito educativo, ¿se asocian los resultados con posibles actividades de instrucción?
	B.	De estar presentes, ¿se incluyen comparaciones grupales de referencia pertinentes acompañadas con información sobre interpretaciones apropiadas?
	C.	De estar presentes, ¿se incluyen los resultados de desempeño en preguntas individuales junto con una clave que explique los atributos del ítem y los códigos de desempeño?
	D.	Si se incluyen datos sobre subescalas, ¿se entrega información a los usuarios sobre el nivel de imprecisión de las puntuaciones? Si se entregan normas, ¿se describe con suficiente detalle el grupo de referencia? ¿Se aclara el significado de términos como puntuaciones T, puntuaciones z, puntuaciones z normalizadas, puntuaciones estaninas, puntuaciones sten, percentiles, puntuaciones equivalentes a grados escolares, etc.?
	E.	Si están presentes, ¿se explican los informes de resultados correspondientes a otros tests recientes y relevantes (por ejemplo, tests con referencia a normas)?
V. Contenido – Otros	A.	¿El informe indica números telefónicos, sitios web o direcciones físicas donde puedan hacerse consultas?
	B.	¿El informe entrega enlaces a recursos adicionales que permitan entender mejor el test, el programa de evaluación y/o el desempeño de quien rinde el test?

VI. Lenguaje	A.	¿El informe está libre de terminología estadística y técnica y de símbolos que puedan confundir a los usuarios?
	B.	¿El texto está claramente escrito para los usuarios?
	C.	¿El informe (o los materiales adicionales) está traducido/adaptado a otros idiomas? De ser así, ¿la traducción fue realizada por más de una persona y se tomaron medidas para validar la versión traducida/adaptada?
VII. Diseño	A.	¿El informe está dividido clara y lógicamente en secciones distintas para mejorar su legibilidad?
	B.	¿Hay una sección de puntos importantes o de resumen que comunique los resultados principales?
	C.	¿El tamaño de letra de las diferentes secciones es adecuado para el público objetivo?
	D.	Si hay gráficos, ¿se presentan claramente al público objetivo?
	E.	¿Se incluye una combinación de texto, tablas y gráficos para apoyar y facilitar la comprensión de los datos y la información del documento?
	F.	¿El informe se ve amigable y atractivo para los usuarios?
	G.	¿El informe transmite una "sensación" moderna mediante el uso efectivo del color y la densidad (una buena tasa de contenido/espacio en blanco)?
	H.	¿El informe está libre de material irrelevante y/o material que podría no ser necesario de acuerdo a los objetivos buscados?
	I.	¿Está claro para el público objetivo el "flujo" de lectura del informe con respecto a dónde sería conveniente empezar a leer?
	J.	¿El informe coincide en cuanto a diagramación y diseño con los materiales relacionados publicados por el programa de evaluación?
VIII. Guías interpretativas y materiales adicionales	A.	¿Hay alguna guía interpretativa preparada? De ser así, ¿es informativa y está escrita de forma clara? ¿Se ha realizado trabajo de campo para probarla? ¿Hay versiones disponibles en varios idiomas para satisfacer las necesidades de los lectores a quienes apunta el informe?
	B.	Si existe una guía interpretativa, ¿se entrega una explicación sobre las interpretaciones aceptables y no aceptables de los resultados del test?

B. Protocolo de grupos focales

Protocolo de grupos focales

1. (2 minutos) Describa brevemente el propósito del proyecto (mejorar una pauta existente haciéndola más accesible, etc.) Haga hincapié en que las críticas son bienvenidas.
2. (2 minutos) Explique las tareas que deberán realizar los participantes (primero, usar la pauta individualmente para evaluar un informe de resultados; luego, tener una conversación grupal sobre su experiencia al usar la pauta).
3. (5 minutos) Revise las instrucciones de la primera página de la pauta junto con los participantes. Consulte si hay preguntas.
4. (1 minuto) Explique cómo se usará la grabación del grupo focal y solicite autorización para grabar la conversación.
5. (Aproximadamente 20 minutos) Revisión individual de la pauta.
6. (Aproximadamente 25 minutos) Conversación grupal.

Preguntas para la conversación grupal

Utilidad

1. ¿Es esta pauta una herramienta útil para evaluar la calidad de un informe de resultados? ¿Por qué? ¿Por qué no?

Estructura

2. ¿Qué les pareció la secuencia de los ítems?
    - ¿Tenía sentido el orden usado?
    - ¿Les pareció que algunos ítems aparecían demasiado pronto o demasiado tarde?
3. ¿Qué escala de evaluación creen que tendría sentido usar con estas preguntas?
4. Considerando todas las preguntas, ¿de qué maneras se podría resumir cuán bueno es un informe de resultados? En otras palabras, después de evaluar un informe de resultados con esta pauta, ¿cómo podríamos entregarle a alguien un "resultado general"?
5. ¿Qué les pareció la forma en que se estaban agrupados los ítems (por ejemplo, "global", "lenguaje")?
    - ¿Podrían recomendar otras formas de agrupar los ítems igual de útiles o tal vez más útiles?

Contenido

6. ¿Hubo preguntas que tuvieran palabras confusas o que en general les parecieran confusas?
    - ¿Cuáles?
    - ¿Cómo mejorarían estas preguntas?
7. ¿Hay preguntas que les hayan parecido menos importantes y que podrían ser eliminadas de la pauta?
8. ¿Hay preguntas que no son parte de la pauta, pero que deberían hacerse cuando se evalúa un informe de resultados?

Habilidades

9. ¿Crees que esta evaluación podría ser respondida por una sola persona? ¿Quién sería esa persona? ¿A qué área del conocimiento debería pertenecer?
    - ¿Qué piensan sobre entregarles distintas partes de esta pauta a grupos de personas diferentes de acuerdo a su área de conocimiento?

Otros

10. ¿Qué otro formato podría ser mejor para presentar estas preguntas?
11. ¿En qué contextos creen que podría usarse esta pauta?

C. Muestra de informe de resultados