Banco Interamericano de Desarrollo

División de Educación (SCL/EDU)

NOTA TÉCNICA

IDB-TN-738

Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles

Catalina Covacevich

Diciembre 2014

Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles

Catalina Covacevich

Banco Interamericano de Desarrollo

2014

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo

Covacevich, Catalina.
Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles / Catalina Covacevich.
p. cm. — (Nota Técnica del BID ; 738)
Incluye referencias bibliográficas.
1. Education. 2. Educational tests and measurements. I. Banco Interamericano de Desarrollo. División de Educación. II. Título. III. Series.
IDB-TN-738

http://www.iadb.org

Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles

Catalina Covacevich[1]

Resumen Ejecutivo

La implementación de prácticas y políticas educativas está muy vinculada a la evaluación de los aprendizajes de los estudiantes, ya que ésta permite monitorear avances y logros, mejorar la enseñanza en el aula, perfeccionar las políticas y evaluar la efectividad de programas, entre muchos otros objetivos. Para que la evaluación logre sus propósitos, es esencial hacer una adecuada elección de los instrumentos de evaluación de aprendizaje que serán utilizados. En este contexto, docentes, implementadores de política, investigadores y personal de los ministerios de educación se ven frecuentemente enfrentados a la necesidad de seleccionar instrumentos de evaluación de aprendizajes, sin necesariamente tener mayores conocimientos sobre el tema. Por lo tanto, esta nota técnica está orientada a personas que trabajan en el sector educación y no tienen formación en el área de evaluación de aprendizajes, y tiene como objetivo entregar orientaciones técnicas, prácticas y éticas sobre los elementos que deben ser tomados en cuenta al seleccionar o construir un instrumento de evaluación de aprendizaje.

JEL: I200, I210, I280
Palabras claves: Educación-Evaluación-instrumentos-pruebas

[1] La autora agradece el apoyo de Daniela Jiménez en la obtención de bibliografía, los valiosos aportes de Hugo Ñopo (SCL/EDU) y los comentarios y guía recibidos durante la preparación del documento de Emiliana Vegas (SCL/EDU), Jefa de la División de Educación.

Contenido

Introducción

El Marco Sectorial de Educación y Desarrollo Infantil Temprano del Banco Interamericano de Desarrollo destaca la necesidad de que los sistemas educativos de aseguramiento de la calidad tengan definidas metas altas de aprendizaje que guíen su quehacer. Tan importante como la definición de estas metas es su monitoreo, el que se puede realizar a distintos niveles del sistema; ya sean evaluaciones de aprendizaje a gran escala, subnacionales, nacionales, regionales o internacionales; o evaluaciones locales, implementadas por los responsables de las escuelas o por los docentes con fines formales, diagnósticos o formativos. Todas estas evaluaciones son insumos para para monitorear el aprendizaje, informar la enseñanza, e incorporar esta información en las actualizaciones de las prácticas y políticas existentes y en el diseño de las nuevas. Por lo tanto, evaluar es parte inherente de un proceso de mejoramiento continuo (Banco Interamericano de Desarrollo, 2013).

Además, cada vez que se implementan programas y políticas educativas, y sobre todo cuando existen recursos limitados, una pregunta crucial es si los programas que se están implementando son efectivos, por lo que la evaluación educativa ha cobrado cada vez más importancia en este contexto. En los últimos años las evaluaciones de impacto, que buscan medir el efecto de un programa comparando grupos equivalentes, en que uno fue beneficiario del programa y otro no, han recibido especial interés.

También se evalúa aprendizajes para la rendición de cuentas, certificación de competencias obtenidas en un determinado ciclo de enseñanza, identificación del nivel de conocimientos de un estudiante para asignarle la formación más apropiada y para selección académica.

La efectividad de todas estas evaluaciones depende en gran parte de la adecuación y calidad de los instrumentos de evaluación de los aprendizajes estudiantiles que se utilicen. Decidir cómo medir el aprendizaje requiere considerar diversos factores, evaluar los pros y contras de cada instrumento disponible, y decidir cuál es el más adecuado para la situación específica. Es poco probable que exista un instrumento ideal, por lo que hay que estar preparado para soluciones de compromiso. Para asegurarse de que se tomó la mejor decisión posible, es necesario asegurarse de haber revisado la mayor cantidad de instrumentos disponibles y de haber recogido la suficiente información sobre cada instrumento (Center for Assessment and Research, James Madison University, 2014).

La presente nota está dirigida a todos aquellos docentes, políticos, diseñadores e implementadores de programas educativos que se ven enfrentados a la necesidad de escoger un instrumento de evaluación de aprendizajes[2] para alguno de los objetivos recién descritos. Tiene

[2] En esta nota de usa el término “aprendizaje” de manera genérica, abarcando habilidades, competencias, dominio de contenidos y cumplimiento de objetivos.

un énfasis en los instrumentos estandarizados a gran escala, pero la mayoría de los principios son aplicables también a instrumentos diseñados y aplicados de manera local. Asimismo, aunque el énfasis es en los instrumentos de evaluación de aprendizajes estudiantiles, muchos de los elementos abordados son aplicables a otros tipos de instrumentos, por ejemplo de evaluación docente.

Algunas de las preguntas que surgen frecuentemente durante el proceso de elección de un instrumento y que serán abordadas en esta nota, son: ¿Para qué queremos evaluar? ¿El instrumento se ajusta a mis necesidades? ¿Qué mide el instrumento? ¿Qué tan útil es este instrumento comparado con otros y en términos de costo y de ahorro de tiempo? ¿Los puntajes son consistentes? ¿Es fácil de administrar? ¿Tiene sesgos culturales o de género? (Cohen y Swerdlik, 2009; Bart, 2009). Una última pregunta muy relevante es si el test necesario está disponible comercialmente o será necesario desarrollar un test propio.

La primera sección de la nota, *Coherencia entre el objetivo de evaluación y el instrumento escogido*, destaca la importancia de que el instrumento seleccionado sea coherente con el objetivo de la evaluación, describe los diversos propósitos evaluativos que existen e identifica las características del instrumento que es necesario contrastar con el objetivo que se tiene en mente. La sección II *Calidad de los instrumentos* aborda la importancia de la calidad técnica de los instrumentos, describiendo las dos principales características técnicas que es necesario tener presentes: la confiabilidad y la validez. La sección III *Consideraciones prácticas* describe temas prácticos importantes de tener en cuenta, tales como el costo del instrumento, que sea de un largo apropiado y fácil de administrar y puntuar. La sección IV *Tipos de instrumentos* presenta diferentes maneras de clasificar los instrumentos de evaluación de aprendizajes, describiendo los distintos tipos de instrumentos que existen según cada clasificación. En la sección V *Consideraciones éticas y justicia de los instrumentos* se discuten una serie de elementos relativos a la ética, el sesgo y la justicia de los instrumentos de evaluación de aprendizajes y cómo estos se relacionan con la validez. La última sección *Crear un instrumento o utilizar uno ya existente* presenta una breve discusión sobre determinar si ya existe un instrumento que sea adecuado a la situación evaluativa, o es más conveniente diseñar un instrumento nuevo, y analiza los pros y contras de ambas situaciones. Por último, se cierra con las conclusiones.

Sabía usted que…

Se cree que los tests y los programas de evaluación surgieron en China en el 2200 AC. Se utilizaban para seleccionar a candidatos que estaban postulando a puestos de gobierno, y consideraban temas tan amplios como música arquería, escritura, matemática, geografía y estrategia militar (Cohen y Swerdlik, 2009).

I Coherencia entre el objetivo de evaluación y el instrumento escogido

Una consideración fundamental para escoger un instrumento de evaluación de aprendizajes es que sea adecuado para el propósito de la evaluación. En esta sección se comienza por describir algunos posibles objetivos de evaluación, para luego abordar los elementos que deben revisarse en un instrumento para ver si son coherentes con el objetivo de evaluación.

A. Para qué queremos evaluar

Determinar el objetivo o propósito de evaluación implica hacerse las preguntas ¿para qué queremos medir aprendizajes?; ¿qué es lo que queremos medir? y ¿a quiénes queremos evaluar? Los posibles motivos para evaluar aprendizajes son muchos. A continuación se identifican algunos de ellos, agrupados en objetivos evaluativos para el sistema educativo en su totalidad, la escuela o el estudiante.

Obtener información a nivel de sistema educativo

Tomar una fotografía de cómo está un sistema educativo. Para poder tomar decisiones de política educativa, un sistema educativo, ya sea a nivel nacional, estatal o municipal, puede necesitar información sobre cómo los estudiantes están logrando los objetivos de aprendizaje propuestos. Para esto, lo más apropiado es diseñar una prueba que evalúe el currículo nacional (o local), en las asignaturas y grados que se consideren más relevantes. En general estas pruebas se aplican al final de los ciclos escolares, para medir los logros esperados para cada ciclo, pero en algunas ocasiones puede ser apropiado tener alguna medición intermedia, para conocer el grado de avance en las metas de aprendizaje y poder intervenir a tiempo, por ejemplo en etapas tempranas de la adquisición de la lectoescritura.

Si solo se desea obtener información agregada de lo que está sucediendo con el conjunto de escuelas o estudiantes, basta con realizar una evaluación muestral, es decir, no es necesario aplicar las pruebas a todas las escuelas ni estudiantes sino solo a una muestra representativa de ellos, lo que hace la evaluación más barata y simplifica los procesos logísticos. Si también se desea información específica a nivel de escuela o de estudiante, y no solo del sistema educativo en su totalidad, la evaluación debe ser censal, es decir, aplicada a todos los estudiantes y escuelas.

En general las pruebas nacionales demoran varios meses en entregar los resultados, por lo que no permiten tener información de manera inmediata, sino al año escolar siguiente.

Comparar con otros países. En ocasiones, más que querer evaluar el desempeño de los estudiantes de un país contra su currículo nacional, se desea tener información de cuánto están aprendiendo los estudiantes del país en comparación con los de otros países. Para esto, un país o subsistema nacional puede participar de los estudios internacionales tales como PISA[3], TIMSS[4],

[3] Programme for International Student Assessment
[4] Trends in International Mathematics and Science Study

PIRLS[5] y las pruebas PERCE, SERCE y TERCE[6] del Laboratorio Latinoamericano de Evaluación de la Calidad Escolar -LLECE-. Todos estos estudios evalúan aprendizajes de estudiantes a nivel escolar en un conjunto de países (o subsistemas nacionales), lo que permite compararlos contra un marco de evaluación conocido y acordado por los países participantes. Cada estudio mide diferentes asignaturas, en grados diferentes y con énfasis diferentes. Por ejemplo, tanto PISA como TIMSS evalúan matemáticas y ciencias, pero mientras TIMSS lo hace en estudiantes de 4º y 8º grado y con un enfoque curricular, PISA lo hace en estudiantes de 15 años, con un enfoque de habilidades para la vida. En el caso de TERCE, solo se evalúan países de Latinoamérica, mientras que en los otros estudios nombrados participan países de diversos continentes.

Estos estudios además recogen información de los estudiantes y los sistemas escolares que permite identificar las variables que más inciden en los aprendizajes, y también sirven para comparar el currículo nacional con el marco de evaluación del estudio, lo que en ocasiones lleva a realizar adaptaciones en los currículos.

Estos estudios son muestrales, ya que se levanta información a nivel del sistema y no de cada escuela, y son conducidos por organismos internacionales de renombre, como La Organización para la Cooperación y el Desarrollo Económicos (OCDE), La Asociación Internacional para la Evaluación del Logro Educativo (IEA) y La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). Son elaborados, administrados y analizados bajo estrictos estándares de calidad que aseguran que los resultados de los países sean comparables.

En general la participación de un país en un estudio internacional toma al menos tres años, ya que el primer año se diseñan y pilotean los instrumentos, al año siguiente se administran y en general recién al año siguiente se liberan las bases de datos y entrega el informe internacional. Ya que no se obtienen resultados inmediatos son útiles para diseño de políticas a mediano y largo plazo, no en el corto plazo.

Medir la evolución de un sistema a lo largo del tiempo. Tanto las evaluaciones nacionales como las internacionales permiten obtener una fotografía de cómo están los logros de aprendizaje en un momento puntual, pero también pueden servir para monitorear el avance en el nivel de aprendizaje a lo largo del tiempo. Por este motivo estas pruebas se administran de manera periódica, en general de manera anual en el caso de las pruebas nacionales, y en ciclos de tres o más años en el caso de las internacionales. Esto permite ir viendo avances en el aprendizaje promedio de cada país y también identificar otras tendencias, como si han disminuido las brechas de género o entre grupos socioeconómicos, o si se evidencian mejoras en ciertas áreas temáticas o en determinados tipos de escuelas después de algunas intervenciones puntuales, como podrían ser cambios en el currículo o intervenciones en ciertos grupos de escuelas. En el caso de los estudios internacionales también se puede comparar el progreso (o falta de progreso) de un país con los avances que otros países han tenido en el mismo periodo.

[5] Progress in International Reading Literacy Study

[6] Primer, Segundo y Tercer Estudio Regional Comparativo y Explicativo

Obtener información a nivel de escuela

En otras ocasiones se desea obtener información de los aprendizajes que se están logrando a nivel de cada escuela. Esto puede deberse a objetivos muy diferentes, tales como retroalimentar al equipo docente y directivo para mejorar los aprendizajes, la rendición de cuentas de quienes son responsables del desempeño de la escuela, informar a los padres y comunidad del desempeño de una escuela, o evaluar el impacto de determinados programas en ciertas escuelas. Muchas veces las evaluaciones a nivel de escuela son conducidas por el gobierno nacional o local, pero también pueden ser ejecutadas por otros organismos, por ejemplo universidades que son contratadas para evaluar la efectividad de algún programa, o incluso por una escuela o grupo de escuelas, que quieren evaluarse a sí mismas para poder monitorear su desempeño y mejorar sus prácticas. Según el objetivo que se tenga en mente, puede ser o no relevante para la evaluación poder realizar comparaciones entre las escuelas evaluadas.

El contenido de las evaluaciones a nivel escuela varía según el objetivo evaluativo: en la mayoría de los casos probablemente se quiera evaluar el currículo nacional, pero en otros puede ser relevante poner el foco con más detalle en algún elemento específico del currículo en que se sabe que hay debilidades, o en el caso de la evaluación de algún programa, puede ser pertinente focalizarse en los elementos específicos que ese programa buscaba promover.

En ocasiones las pruebas nacionales están diseñadas para entregar información por escuela, en otras puede ser necesario diseñar pruebas especiales, ya sea porque esta información no está disponible o porque el foco de la evaluación nacional no es el más adecuado para el objetivo de evaluación.

Obtener información a nivel de estudiante

En algunos casos se busca obtener información a nivel del estudiante. Esto puede darse como parte del proceso de mejoramiento continuo del sistema, para poder identificar las debilidades de un determinado estudiante y que sus profesores y padres lo puedan apoyar. En otros contextos se aplican pruebas que certifican los aprendizajes logrados, por ejemplo pruebas de fin de ciclo, como podría ser el caso de una licencia de secundaria o un examen profesional, o una prueba que certifique un determinado nivel de dominio de un idioma extranjero. Los instrumentos de aprendizaje individuales también pueden ser utilizados para seleccionar estudiantes, por ejemplo las pruebas de selección universitaria. En otras ocasiones, se evalúa a los estudiantes al comienzo de un año escolar o programa educativo, para tener un diagnóstico de sus debilidades y fortalezas y poder reforzar los elementos necesarios o hacer cursos de nivelación.

B. Alineación entre los objetivos de evaluación y el instrumento

Un elemento fundamental al momento de escoger un instrumento de evaluación de aprendizajes es que sea adecuado para el o los objetivos de la evaluación. Este alineamiento es crucial porque se relaciona con la utilidad de la información que se obtendrá. Si el alineamiento es bajo, los resultados de la evaluación entregarán poca o limitada información (Center for Assessment and

Research, James Madison University, 2014). Para esto, se debe revisar el objetivo, contenido y población objetivo declarados por el instrumento y asegurarse que se alinean con el propósito de evaluación.

Alineación de los propósitos

Una primera pregunta es para qué propósito fue diseñado el test, y si calza con el propósito de evaluación (Ministerio de Educación de Nueva Zelandia, 2014). Los propósitos de un instrumento pueden ser varios, tales como realizar un diagnóstico, medir logro, medir potencial o aptitud, o identificar preparación para un cierto programa o etapa escolar (llamado placement testing en inglés), que también se pueden utilizar para ubicar a un estudiante en un cierto programa o track de aprendizaje (Cohen, Manion y Morrison, 2000).

Algunos instrumentos están diseñados para realizar evaluaciones diagnósticas, formativas, o sumativas. La evaluación diagnóstica es una evaluación en profundidad en las debilidades y fortalezas de un estudiante. En general incluye muchos ítems que profundizan en un solo tema, para poder identificar con exactitud las dificultades de aprendizaje, y es referida a criterios[7]. La evaluación formativa, en cambio, ocurre durante un año escolar o programa y está diseñada para para monitorear el progreso del estudiante durante ese periodo, para medir logros de secciones específicas del currículo, para diagnosticar debilidades y fortalezas. En general está referida a criterios. La evaluación sumativa se aplica al final del curso o programa, y está diseñada para medir logros o outcomes. Puede ser referida a normas o a criterios, dependiendo hasta cierto punto del uso que se le dará a la evaluación (por ejemplo, entregar certificados o grados) (Cohen, Manion y Morrison, 2000). Por lo tanto, si el objetivo de evaluación es realizar un diagnóstico, probablemente un instrumento diseñado como evaluación formativa no sea lo más apropiado.

Alineación de los contenidos

Una segunda pregunta es qué es lo que el instrumento mide y si esto calza con lo que se quiere evaluar. Este análisis no es algo general, como por ejemplo, "habilidades matemáticas en primaria". Hay que mirar en detalle el contenido del test, los niveles de aprendizaje que cubre, y el o los grupos etarios a los que está orientado, y compararlo minuciosamente con los objetivos de evaluación. Por ejemplo, para una evaluación el propósito puede ser evaluar la implementación del currículo nacional de matemática, el que puede o no tener elementos en común con una prueba de matemática diseñada para su uso universal, ya que el currículo puede tener ver con medir aprendizajes de geometría, mientras que un test puede tenerlo en aritmética. Es muy importante analizar en detalle el contenido de la prueba, y ver si cubre adecuadamente los contenidos que se quieren evaluar, y además comprobar que no evalúe elementos que no son parte del objetivo de evaluación. Este análisis se vuelve a abordar en la sección referida a los

[7] En la sección "Tipos de instrumentos" se explica la diferencia entre la evaluación referida a criterios y aquella referida a normas.

argumentos sobre la evidencia basada en el contenido, en la sección sobre la validez del instrumento.

También es necesario mirar en detalle el contenido del instrumento para asegurarse de que sea adecuado al nivel cognitivo, de lectura, y otras habilidades, de los evaluados. Por ejemplo, pruebas que requieran un nivel lector avanzado no pueden ser administrado a usuarios con poca comprensión lectora (Timmons et al, 2005).

Alineación de la población objetivo

Un tercer elemento a tomar en cuenta es la coherencia entre la población para la que fue diseñado el instrumento y la que se desea evaluar. En el caso de las pruebas relativas a normas, también hay que revisar para qué población fueron construidas estas normas (Ministerio de Educación de Nueva Zelandia, 2014).

En los comienzos de la evaluación, aunque muchos de los instrumentos publicados estaban diseñados expresamente para una población específica, estos se administraban -inapropiadamente- a personas de diferentes culturas. Y, de manera no sorprendente, los evaluados pertenecientes a minorías tendían a obtener puntajes más bajos que aquellos para quienes fue desarrollado el instrumento. Por ejemplo, un ítem del test WISC de 1949 preguntaba: "si tu madre te manda al almacén por una barra de pan, y no hay ninguna, ¿qué haces?" Muchos niños latinos iban seguido a comprar tortillas, pero no pan, y por lo tanto no entendían la pregunta ni sabían qué responder (Cohen y Swerdlik, 2009).

Hoy en día, los desarrolladores de instrumentos toman muchas precauciones para que estos sean adecuados para la población para la que fueron diseñados. Por ejemplo, que una prueba diseñada para uso nacional, efectivamente sea apropiado para la población nacional. Estas precauciones pueden incluir: administrar una versión piloto a estudiantes de diferentes características; preguntarle sus impresiones a los examinadores, por ejemplo, sus opiniones subjetivas sobre la calidad de las instrucciones; analizar los ítems para ver si presentan sesgo racial, cultural, o de género; o pedirle a un panel de expertos que revisen los ítems buscando posible sesgo (Cohen y Swerdlik, 2009). También los estudiantes con necesidades educativas especiales pueden presentar dificultades, lo que se aborda en la sección referida a ética.

A veces existen versiones adaptadas de instrumentos que originalmente fueron diseñados para otra población. Muchas veces en países latinoamericanos se utilizan versiones adaptadas de instrumentos diseñados en Estados Unidos. Estas adaptaciones incluyen la construcción de normas para la población de ese país específico y pueden haber sido realizadas por los mismos diseñadores del test o en otras ocasiones, por universidades locales u otras instituciones dedicadas a la investigación. En ocasiones, es posible basarse en la versión adaptada para un país vecino, pero hay que revisar exhaustivamente los ítems y hacer una prueba piloto para ver si las adaptaciones son adecuadas para la población específica que se requiere evaluar.

II. Calidad de los instrumentos

Otro importante aspecto que surge al analizar un instrumento es su calidad técnica. Cuanto mejor es la calidad de un instrumento, más útil será, más confianza se le puede tener a los puntajes obtenidos y mayor será la confianza para tomar decisiones a partir de estos resultados, por lo que es imperativo usar instrumentos de alta calidad al hacer evaluaciones (Center for Assessment and Research, James Madison University, 2014). Los dos principales elementos que dan cuenta de la calidad de un instrumento son su validez y su confiabilidad.

A Validez

Evolución del concepto validez

El concepto de validez ha ido sufriendo transformaciones a lo largo del tiempo. Tradicionalmente, la validez de un instrumento se ha entendido como hasta qué punto el instrumento efectivamente mide lo que sus autores declaran que mide (Cohen, Manion y Morrison 2000; Darr, 2005). Sin embargo, más recientemente los especialistas en evaluación han considerado que la validez no es una propiedad fija e inherente del instrumento, sino que es un juicio, basado en evidencia, sobre qué tan apropiadas son las inferencias realizadas o acciones implementadas a partir de los puntajes de una prueba en un determinado contexto (Salvia & Ysseldyke, 2004; Cohen y Swerdlik, 2009). Entonces, la validación puede ser vista como el desarrollo de un argumento de validez sólido para los usos propuestos de los puntajes de un instrumento y su relevancia para el uso propuesto (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 1999)[8]. Definida de esta manera, la validez no es una propiedad inherente al instrumento, sino que se relaciona con el objetivo de evaluación.

Ya que lo que se juzga no es en realidad la validez del instrumento ni de sus puntajes, sino la de la interpretación de los resultados del test para determinados usos, cuando se pretende usar un test de varias maneras, la validez de cada uso se debe analizar de forma separada (Joint Committee on Standards for Educational and Psychological Testing, 1999).

En consecuencia, si un test es válido para un determinado grupo o población, no necesariamente lo será para otros. Por ejemplo, si un test de razonamiento espacial diseñado para estudiantes con un nivel de lectura de 8º grado, es aplicado a un estudiante con habilidad lectora de 4º grado, sus resultados pueden reflejar tanto su nivel lector como su capacidad de razonamiento espacial.

[8] Se ha llamado argumento interpretativo a la declaración explícita de las inferencias y supuestos a la base de los usos e interpretaciones propuestas. Y argumento de la validez, a la evaluación de la coherencia del argumento interpretativo y de la plausibilidad de sus inferencias y usos. Por lo tanto, es necesario que la interpretación y los usos deben estar explicitados (Brennan, 2006).

Esta nueva manera de entender la validez está muy relacionada con la evaluación de programas, en que es necesario especificar el programa que se evaluará, los contextos en que se implementará, se deben descartar variables externas que puedan afectar los resultados, y porque muchas veces los programas se evalúan más con un conjunto de evaluaciones que con un estudio aislado. De la misma manera, se puede entender la validez como una evaluación integral de los usos propuestos para la evaluación, generando un análisis coherente de toda la evidencia a favor y en de dicho uso y si es posible, sobre explicaciones alternativas (Cohen y Wollak, 2006).

Algunas preguntas relativas a la validez de un test se cuestionan la calidad de sus ítems: ¿Los ítems son una muestra adecuada del constructo que se quiere evaluar? También hay preguntas relativas a la interpretación de los resultados: ¿Qué nos dicen estos puntajes? ¿Cómo se relacionan los puntajes altos o bajos con el comportamiento de los evaluados? ¿Cómo se relacionan estos puntajes con los de otros instrumentos que dicen medir lo mismo? (Cohen y Swerdlik, 2009).

La validez debería estar como prioridad al diseñar o seleccionar instrumentos para la evaluación de aprendizajes. Es crítico que la evaluación permita hacer juicios sobre los progresos de los estudiantes que sean robustos y útiles, y tengan consecuencias positivas. Estar consciente de la validez y de cómo esta se puede ver amenazada puede ayudar a tomar decisiones sobre qué evaluaciones vale la pena hacer y qué usos se les puede dar a estas evaluaciones (Darr, 2005).

¿Quién es responsable por la validez?

Es responsabilidad del desarrollador del test entregar evidencia sobre la validez de su instrumento, especificando la población en la que fue validado. Pero es responsabilidad del usuario evaluar si el instrumento es apropiado al contexto particular en que lo aplicará. En ocasiones, puede ser apropiado que el usuario conduzca estudios extras de validación local. Esta validación local se vuelve imprescindible cuando se planea hacer alguna modificación al instrumento en sus instrucciones, idioma del instrumento, o si se pretende aplicarlo a una población que sea significativamente diferente a aquella en que el test fue estandarizado, o si se le quiere dar un uso diferente de aquello para que fue diseñado (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009).

¿Cómo se mide la validez?

Juzgar si un instrumento es válido no es algo que pueda ser medido en una escala absoluta. Frecuentemente se categoriza la validez como débil versus aceptable lo que refleja un juicio sobre qué tan adecuadamente el test mide lo que se supone que mide (Cohen y Swerdlik, 2009). Otros autores, como Darr (2005), sugieren que se categorice como débil, moderada, o fuerte.

Ya que la validez está referida a inferencias y decisiones hechas para un grupo específico en un contexto específico, para juzgar la validez de un instrumento se requiere reunir mucha información (Darr, 2005). Existen diferentes enfoques, orientados a probar diferentes tipos de validez. Estos enfoques no son mutuamente excluyentes, y todos contribuyen a la validez total, aunque según el uso que se le quiera dar al test pueden tener distinta relevancia (Cohen y Swerdlik, 2009). Clásicamente, se habla de validez de constructo, contenido y criterio. Distintos autores hacen distintas clasificaciones de los tipos de validez que se deben considerar, por ejemplo validez de constructo, contenido, ítem, predictiva, "face", relativa a criterios, concurrente, etc, aclarando que no es necesario usar siempre todas las formas de validez (Wilson, 2005).

En este documento se utilizará el enfoque propuesto en los Estándares para la Evaluación Educativa y Psicológica (Joint Committee on Standards for Educational and Psychological Testing, 1999), en que más que hablar de tipos de validez, se habla de *tipos de evidencia* sobre la validez, o líneas de evidencia, basadas en el contenido del test, los procesos de respuesta, la estructura interna, las relaciones con otras variables, y las consecuencias (Joint Committee on Standards for Educational and Psychological Testing, 1999; Wilson, 2005). Estos tipos de evidencia se describen en el siguiente cuadro.

Cuadro 1. Tipos de evidencia sobre la validez

Evidencia basada en el contenido	En los tests no se pueden evaluar todos los conocimientos de los estudiantes, sino solo una muestra de ellos, por lo tanto es muy importante que esta sea una muestra adecuada del área de aprendizaje que interesa evaluar. Si esto se logra, aumenta nuestra posibilidad de hacer inferencias válidas sobre los logros de aprendizaje en un cierto dominio (Darr, 2005). Este tipo de evidencia requiere mirar el contenido del instrumento para analizar la relación con el constructo que se quiere medir (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009). Para saber qué es lo que un test mide, no basta con guiarse por el nombre, es fundamental mirar los ítems que lo componen (Center for Assessment and Research, James Madison University, 2014). Se puede analizar cada ítem en relación con el dominio, o pedir la opinión de expertos sobre qué tan adecuadamente estos abordan el dominio (Joint Committee on Standards for Educational and Psychological Testing, 1999). Si un instrumento es bueno, tendrá ítems que evalúen diferentes aspectos del tema evaluado, y expertos en el área, que no están familiarizados de antemano con los ítems, estarán de acuerdo en qué evalúa cada ítem. Existen dos riesgos que deben ser evitados. Uno es la sub-representación del constructo, es decir, que elementos importante del constructo que se quiere evaluar no estén siendo evaluados. El otro es la varianza relacionada con constructos que son irrelevantes para lo que se está midiendo, por ejemplo en una prueba de lectura, el conocimiento previo del tema o la respuesta emocional frente al texto, o en un test de matemática, la velocidad de lectura o el vocabulario (Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidencia basada en los procesos de respuesta	Los análisis teóricos y empíricos sobre los procesos de repuesta de los examinados pueden entregar información sobre la relación entre estos procesos y los constructos que se desean evaluar. Por ejemplo, si un test busca evaluar razonamiento matemático, es importante que el test efectivamente evalúe eso y no simplemente la aplicación de algoritmos. Observar estrategias de respuesta o entrevistar a los examinados sobre los procesos puede entregar esta información (Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidencia basada en la estructura interna	Este análisis busca recoger evidencia sobre el grado en que las relaciones entre los ítems de un test y sus componentes se adecúan al constructo que supuestamente buscan evaluar, el que puede implicar una sola dimensión, o varias. Para mirar esto se puede revisar si los ítems efectivamente cumplen con el mapa de contenidos. Si el constructo tiene una sola dimensión, esto también se puede probar a través del análisis de ítems (por ejemplo, que a los estudiantes con un buen desempeño en el total de la prueba, obtengan un buen desempeño en el ítem). Otra forma de mirarlo es comprobar que los ítems funcionen de manera diferente en distintos grupos, de acuerdo a lo que predice la teoría (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Cuadro 1. Tipos de evidencia sobre la validez (continuado)

Evidencia basada en relaciones con otras variables[9]	Este tipo de evidencia se desglosa en validez convergente[10] y discriminatoria. La evidencia relativa a la validez convergente implica comparar los resultados obtenidos en un determinado test con los obtenidos por los mismos estudiantes en tests que midan el mismo constructo, o constructos similares. Se espera que los puntajes de un cierto instrumento se correlacionen con otros que declaran medir constructos iguales o parecidos (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999): si dos evaluaciones que supuestamente miden el mismo constructo están entregando resultados muy diferentes, es motivo de preocupación (Darr, 2005). Una posible dificultad es que muchas veces no existen otros instrumentos parecidos (Wilson; Joint Committee on Standards for Educational and Psychological Testing, 1999). La evidencia relativa a la validez discriminatoria se obtiene comparando los resultados obtenidos en el test con otras evaluaciones que midan constructos opuestos o diferentes. En este caso, se espera que los puntajes se correlacionen poco con los de tests que declaran medir constructos diferentes (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999).
Evidencia basada en las consecuencias	Más allá de toda la información técnica recogida, si el uso de una evaluación en particular tiene o puede tener consecuencias negativas, o las consecuencias de usar sus resultados pueden ir en contra del objetivo educativo final, es una consideración que debe tomarse en cuenta para cuestionarse la validez de un instrumento y decidir si usarlo o no (Darr, 2005; Wilson 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999). Esta es la validez desde el punto de vista de las consecuencias de usar los resultados de los tests. Por ejemplo, el peso que se le dé a los resultados puede tener un impacto en las maneras de enseñar y aprender. Algunas de las consecuencias negativas pueden ser estrechamiento curricular, "teaching to the test" o reducción en la motivación de los estudiantes (Darr, 2005). Para analizar este tipo de evidencia es necesario considerar tanto los efectos intencionados como los no intencionados de los tests (Wilson 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999). También es necesario analizar si las consecuencias indeseables se deben al constructo que se quiere medir, o al instrumento específico que se está utilizando para pedirlo. Para poder dilucidar esto se debe ver si otro instrumento que mida el mismo constructo presenta las mismas consecuencias indeseables. Si este es el caso, es más probable que el problema se deba al constructo que al instrumento (Wilson, 2005). Por último, hay que distinguir las consecuencias que tienen que ver con decisiones de política educativa, pero no necesariamente con la validez. En general, la evidencia relacionada con consecuencias se relaciona directamente con la validez si tiene que ver con la sub representación de un constructo o con la irrelevancia de constructo descritas anteriormente (Wilson, 2005; Joint Committee on Standards for Educational and Psychological Testing, 1999).

[9] Algunos autores identifica este tipo de validez como externa.

[10] En el enfoque tradicional, se habla de validez concurrente, que es el grado en que el puntaje de un test se relaciona con otra medida obtenida al mismo tiempo, es decir, cuanto se relacionan los puntajes de un test que declara medir "x" con los de otro test que dice medir lo mismo. Y de la validez predictiva, que refiere al grado en que el puntaje en un test predice alguna conducta relativa al criterio medida por otro test en el futuro (Cohen−Swerdlik, 2009; Darr, 2005).Por ejemplo, a nivel de estudios secundarios o terciarios, una evaluación predictiva podría tener sentido para ver qué tan bien las evaluaciones a nivel escolar predicen el desempeño académico o laboral futuro (Darr, 2005).

B Confiabilidad

¿Qué es la confiabilidad?

La confiabilidad se refiere a la consistencia con que el instrumento mide, o visto de otro modo, al grado de error presente en la medida (Cohen y Swerdlik, 2009). Un test confiable entrega resultados consistentes a lo largo del tiempo. Por ejemplo, estudiantes con el mismo nivel de comprensión lectora que toman un test de compresión lectora tendrán puntajes similares o idénticos, sin importar cuándo lo tomen, asumiendo que su nivel de comprensión lectora no ha variado (Timmons, Podmostko, Bremer, Lavin y Wills, 2005).

En teoría, un instrumento perfectamente confiable mide siempre de la misma manera (Cohen y Swerdlik, 2009). Por ejemplo, un termómetro. Sin embargo, en realidad la evaluación educativa nunca está libre de algún grado de error, ya que un mismo individuo no siempre rinde de la misma manera y las condiciones externas también pueden inducir a error (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Para mirar la confiabilidad, al igual que la validez, hay que entenderla en contextos y propósitos evaluativos específicos. Sin embargo, ya que la confiabilidad remite a cuánta variación es esperable entre una medición y otra, se entiende de una manera más estrechamente estadística que la validez, que refiere a la naturaleza de los atributos siendo medidos[11] (Haertel, 2006).

¿Quién es responsable por la confiabilidad?

Los sitios web o los manuales de los instrumentos deben especificar su confiablidad. Si no lo hacen sus resultados deben tomarse con mucha cautela y no usarse para tomar decisiones de alto impacto (Timmons et al, 2005).

La información que debe presentarse es la identificación de las principales fuentes de error, resúmenes estadísticos que cuantifiquen el tamaño de estos errores, y el grado de generalizabilidad de los puntajes entre distintas formas, puntuadores, administradores, y otras dimensiones relevantes. También una descripción de la población con el que fueron hechas estas estimaciones. Es necesario que haya bastante detalle para juzgar si la confiabilidad es adecuada, ya que no existe un índice único, aplicable a cualquier situación (Joint Committee on Standards for Educational and Psychological Testing, 1999)[12].

---

[11] Ya que este documento está dirigido a personas que no necesariamente tienen conocimientos de estadística o psicometría, solo se describirán los métodos para estimar la confiabilidad de manera muy general. Para profundizar en el tema, se sugiere consultar "Reliability" de Haertel, E. y "Item Response Theory", de Yen, W. & Fitzpatrick, A., ambos en el libro Educational Measurement, (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

[12] Es importante ver si los análisis fueron hechos con puntajes brutos o no.

¿Cómo se mide la confiabilidad?

Se han desarrollado varios marcos teóricos estadísticos importantes para analizar la confiabilidad. Los principales son la teoría clásica de medición, la teoría de generalizabilidad, y la teoría de respuesta al ítem (TRI) (Haertel, 2006).

Según cuál de estos enfoques se utilice, la confiabilidad se calcula de maneras distintas y también la información puede ser reportada de diversas maneras: como varianza o desviaciones estándares de errores de medición, como uno o más coeficientes, o como funciones de TRI (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009). Estos diferentes enfoques se describen en el siguiente cuadro.

Cuadro 2. Marcos teóricos para analizar la confiabilidad

Teoría clásica	Desde la teoría clásica, los enfoques más típicos para analizar la confiabilidad son: coeficientes derivados de la administración de formas paralelas en sesiones independientes, coeficientes obtenidos por la administración del mismo instrumento en ocasiones separadas (conocido también como "test re-test", o "coeficiente de estabilidad") y coeficientes basados en la relación entre puntajes derivados de ítems individuales o subtests dentro de un test, información que es obtenida de la misma administración (conocido también como "coeficiente interno", o "inter ítem") (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009). El coeficiente de confiabilidad más usado en teoría clásica es el Alpha de Cronbach, que pertenece a esta última categoría. Alpha se desarrolló en 1951 para entregar una medida de la de la consistencia interna de un test o una escala, es decir, identificar cuánto los ítems miden el mismo concepto, por lo tanto si un test tiene varias escalas puede ser más apropiado usar alpha en forma separada para cada escala. Si los ítems se correlacionan entre sí, al valor de alpha aumenta. Pero este valor puede aumentar también por la cantidad de ítems (Webb, Shavelson & and Haertel, 2006). Sus valores posibles se mueven entre 0 y 1. En general un alpha de .7 o más se considera aceptable (Institute for Digital Research and Education, UCLA, n.d.), por ejemplo para la evaluación de programas (Center for Assessment and Research, James Madison University, 2014), pero si los resultados tendrán consecuencias individuales es mejor obtener valores superiores a .8 (Webb, Shavelson & and Haertel, 2006).
Teoría de generalizabilidad	La teoría clásica asume que el puntaje observado es la suma del puntaje verdadero y algún error residual específico de ese puntaje. En cambio, la teoría de generalizabilidad en vez de usar el puntaje verdadero, asume un universo de generalización compuesto por todas las posibles observaciones consideradas equivalentes (Brenan, 2006, Haertel, 2006). Los coeficientes utilizados por la teoría de generalizabilidad permiten especificar y estimar los diversos componentes de la verdadera varianza del puntaje, la varianza del error, y varianza del puntaje observado (Joint Committee on Standards for Educational and Psychological Testing, 1999). Se pueden realizar dos tipos de estudios, de generalizabilidad (G-Study) y de decisión (D-Study). Una herramienta de análisis habitualmente utilizada es ANOVA, así como el programa computacional GENOVA.

Cuadro 2. Marcos teóricos para analizar la confiabilidad (cotninuado)

Teoría de respuesta al ítem	La TRI es una familia de modelos estadísticos usados para analizar los datos de ítems de tests, entregando un proceso estadístico unificado para estimar características de los ítems y los individuos examinados y definir cómo estas características interactúan en el desempeño en los ítems y el test. IRT tiene muchos posibles usos en evaluación, entre ellos construcción de ítems, escalamiento, equating, estándar setting, y puntuación. A partir de los '90 ha sido utilizada en la mayoría de las evaluaciones estudiantiles a gran escala. Existen diferentes modelos TRI pero su esencia común es una descripción estadística de la probabilidad de que un examinado con determinadas características tenga una determinada respuesta a un ítem individual, que a su vez tiene características particulares. Las maneras de calcular la confiabilidad bajo TRI toman en cuenta las características del individuo y de los ítems (Yen & Fitzpatrick, 2006). Al usar TRI muchas veces se utiliza la función de información del test como medida de confiabilidad. Esta resume qué tan bien el test discrimina entre individuos de diversos niveles en el rasgo siendo evaluado (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Estos tres enfoques se refieren a la confiabilidad del instrumento, pero las fuentes de varianza en la medición también pueden darse en la puntuación e interpretación de los instrumentos. Por ejemplo, cuando el proceso de puntuación requiere mucha participación de puntuadores (lo que sucede en los ítems de respuesta abierta), en general se obtienen puntajes de consistencia entre jueces, que es otra forma de analizar la confiablidad (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009).

C Estandarización y su importancia para la validez y confiabilidad

Hay que tener en cuenta que cómo se implemente y analice un instrumento también puede afectar su validez y confiabilidad (Joint Committee on Standards for Educational and Psychological Testing, 1999; Cohen y Swerdlik, 2009). Para que un test sea válido, no basta con que las características técnicas del instrumento lo sean, también es fundamental que todos los instrumentos hayan sido administrados bajo las mismas condiciones estandarizadas de aplicación. Esto significa que las instrucciones, el contexto de aplicación y los procedimientos de puntuación han sido exactamente los mismos para todos los examinados. Eso asegura que los datos puedan ser adecuadamente, interpretados, comparados, y usados de acuerdo al principio de que cada usuario fue tratado de manera justa. Por lo tanto, cualquier alteración en la estandarización de la aplicación afecta la comparabilidad y la validez de la prueba (McCallin, 2006).

Esto nos lleva al concepto estandarización. Lo que hace que un test sea estandarizado no es el uso de puntajes estandarizados, o que sea de respuesta múltiple, sino que las condiciones de aplicación hayan sido estandarizadas, es decir, las instrucciones, condiciones de administración, y puntuación son claramente definidas y son las mismas para todos los examinados (Ministerio

de Educación de Nueva Zelandia, 2014; Cohen y Wollak, 2006). La estandarización es importante para todo tipo de instrumentos, sin importar si son referidos a normas o criterios, del formato que tengan y de si tienen o no distintas formas. Las instrucciones estandarizadas aseguran que todos los examinados tengan el mismo entendimiento de lo que se espera de ellos (Cohen y Wollak, 2006). Ejemplos de alteración de la situación de administración son: examinadores que dan más o menos tiempo para responder los instrumentos; que no leen las instrucciones (llevando a confusión sobre cómo responder o sobre las condiciones de puntuación, por ejemplo si las respuestas malas se descuentan de las buenas); cualquier falta al protocolo de aplicación, por ejemplo un examinador que lee en voz alta las preguntas de comprensión lectora; alterar las instrucciones (tiempos, instrucciones, formato de respuesta, soplar); variabilidad en los centros de aplicación (posters con información relativa a los contenidos del test en la pared; interrupciones durante las sesiones; diferencias en la hora de aplicación, que llevan tener niños con distinto nivel de cansancio y hambre); diferentes condiciones en la temperatura e iluminación; y problemas técnicos relevantes a la aplicación (Cohen y Wollak, 2006; McCallin, 2006). También son faltas a la estandarización si hay presión hacia los estudiantes de rendir bien, si tienen exposición previa a los ítems, o el examinador da algunas respuestas (Cohen y Wollak, 2006; McCallin, 2006).

En la sección referida a ética se describen las maneras en que faltas éticas por parte de evaluados o los responsables de la evaluación pueden afectar la validez de un instrumento, abordando los temas de la intención de engaño, fraude, y trampa. También se profundiza la discusión sobre la relación entre justicia de un instrumento y validez.

III. Consideraciones prácticas

Además de las consideraciones técnicas ya descritas, existen una serie de consideraciones prácticas que también es importante tomar en cuenta al momento de decidirse por un instrumento. Esta sección describe elementos como los costos, tiempos de aplicación, o el entrenamiento que se requiere para los examinadores, que pueden hacer inviable la utilización de un cierto instrumento en determinado contexto, aunque su contenido sea muy apropiado y su calidad técnica excelente. En algunas ocasiones, quizás estos elementos no sean una limitación, pero sí factores que se deben tomar en cuenta en la planificación de la aplicación y en el desarrollo del presupuesto. Por ejemplo, muchas veces se subestima el tiempo y los recursos que requieren el reclutamiento, selección y capacitación de examinadores, la corrección de pruebas de preguntas abiertas, o el tiempo y cantidad de personas necesarias para la digitación o el escaneo de los datos.

Costos

Un elemento central en este análisis es el tema costos, que puede ser un factor decisivo a la hora de escoger un instrumento. Para cada instrumento potencial hay que saber cuántos recursos se necesita tener para implementar la evaluación en su totalidad (Center for Assessment and Research, James Madison University, 2014; Ministerio de Educación de Nueva Zelandia, 2014; Timmons et al, 2005). Existen varios tipos de costos asociados a los instrumentos: para comprar los derechos de los tests mismos, las hojas de respuesta, el procesamiento, y puntuación y análisis de los datos por parte del dueño del test o algún proveedor independiente. Además hay costos asociados al pago de sueldos al personal que administre y puntúe, costos legales o de licencias de estas contrataciones, y el arriendo de un lugar para la evaluación, almacenaje de material y corrección de preguntas abiertas, si corresponde (Cohen y Swerdlik, 2009).

Con respecto a los derechos de los instrumentos, algunos están disponibles sin costo, pero otros deben ser comprados a sus autores o publicadores (Center for Assessment and Research, James Madison University, 2014). Los instrumentos por los cuales hay que pagar tienen un amplio rango de precios y es importante considerar que el mejor instrumento no es necesariamente el más caro. La mayoría de los publicadores de tests de lápiz y papel cobran por manuales y otros materiales de administración, además de por cada test individual, hojas de respuestas, y servicios de puntuación. Si se quiere hacer un análisis de costo-efectividad hay una serie de factores que deben considerarse. Por ejemplo, algunos tests baratos pueden ser útiles o instrumentos muy caros tener una utilidad muy acotada a determinada población. También es importante considerar cuántas veces se planea usar el instrumento y si es posible asociarse con otra institución para compartir costos (Timmons et al, 2005).

Tiempos de aplicación

Los manuales de los instrumentos siempre especifican el tiempo (o los rangos de tiempo) de su aplicación. Esto puede determinar la adecuación del instrumento para un cierto uso (Ministerio de Educación de Nueva Zelandia, 2014). Por ejemplo, puede que el programa no cuente con suficiente tiempo y dinero para una administración larga. En niños muy pequeños, los exámenes muy largos pueden ponerlos ansiosos y hacer los puntajes menos válidos que los de exámenes más cortos. En otros casos, la fatiga puede ser un factor que influya en los resultados. Es importante escoger instrumentos cuya validez no se vea afectada por este tipo de variables (Timmons et al, 2005).

Otra consideración práctica es que a veces las evaluaciones deben ajustarse a la duración de los bloques escolares. Por ejemplo, si hay bloques de 90 minutos entre recreos, quizás es logísticamente complejo aplicar un instrumento que dure 120 minutos (Bart, 2009).

Por lo tanto, hay que evaluar los tiempos de aplicación contra el tiempo disponible y el tipo de estudiante al que se le aplicará el test, para evaluar si es adecuado o no al contexto particular.

Entrenamiento requerido para los examinadores

Los administradores de los instrumentos juegan un rol esencial. Los tests varían en el nivel de experticia y entrenamiento requeridos por los administradores o puntuadores (Ministerio de Educación de Nueva Zelandia, 2014; Timmons et al, 2005).Cuando se requiere experiencia o formación específica, esto se especifica en los manuales o los sitios webs de los instrumentos. En algunas ocasiones, incluso es necesario enviar documentación que respalde la formación de los examinadores antes de poder acceder a los tests. La administración o puntuación de tests por parte de personal sin las calificaciones necesarias es una serie violación ética y además puede afectar la validez de los resultados (Timmons et al, 2005).

Para algunos instrumentos se requiere una capacitación específica para la administración del test (más allá de que se requiera o no una cierta formación previa). Es fundamental tener esta información de antemano, y en el caso de exigirse un entrenamiento, saber cuánto cuesta y cuánto tiempo toma (Ontario Ministry of Training, Colleges and Universities, 2011). También es importante saber si se cuenta con entrenadores adecuados para poder realizar las capitaciones.

Facilidad/dificultad en la puntuación y análisis

Algunos instrumentos requieren tiempo y/o un entrenamiento especial para ser puntuados o para analizar los datos, lo que puede involucrar contratar y entrenar a personas para que lo hagan. Esto es especialmente común en aquellos que evalúan escritura. Es necesario considerar el tiempo necesario y los costos asociados a la selección y capacitación de correctores, lo que

incluye diseñar e implementar un sistema para asegurar la confiabilidad entre correctores, además del tiempo que se dedica a la puntuación (Cohen y Wollak, 2006). También es necesario tomar en cuenta el espacio físico requerido para realizar la puntuación.

Hay que tomar en cuenta también cómo se registrarán los datos y/o puntajes en una base de datos (Ministerio de Educación de Nueva Zelandia, 2014; Center for Assessment and Research, James Madison University, 2014). La puntuación de los tests es más eficiente que antes, en muchos casos con acceso a puntuaciones computarizadas o por internet. En algunas situaciones se puede acceder a la puntuación de manera inmediata (Timmons et al, 2005), pero en otros casos esa información debe ser cargada. La mayoría de los tests hoy se leen con un lector óptico, lo que es mucho más rápido que la digitación. Para esto hay que tener las hojas de respuestas apropiadas y el lector óptico adecuado (Cohen y Wollak, 2006).

Administración grupal o individual

Algunos instrumentos requieren una administración uno a uno (un examinador por evaluado) mientras otros son de aplicación grupal. En términos prácticos, lo más fácil es ocupar instrumentos de aplicación grupal, es decir, que implican tener solo uno o dos examinadores por sala.

Sin embargo, muchos instrumentos que requieren observación por parte del examinador, o que evalúan a niños muy pequeños, son de administración individual, es decir, uno a uno entre el examinador y el examinado. Esto puede ser inviable desde el punto de vista práctico, por costos y limitaciones de tiempo. En otras ocasiones quizás lo que se quiere medir solo se puede evaluar a través de un instrumento de administración individual, en este caso es necesario contar con los recursos suficientes. La mayoría de los instrumentos que han sido diseñados para administración individual no se pueden usar fácilmente en un grupo (Ministerio de Educación de Nueva Zelandia, 2014).

Es fácil/difícil de usar

Los test deberían ser lo más fáciles de usar posible, ya que los desempeños de los estudiantes se pueden ver afectados si no entendieron las instrucciones. Sobre todo los niños pequeños pueden gastar tiempo valioso tratando de entender el proceso y no en el contenido. Por ejemplo, hojas de respuesta complejas pueden confundir al usuario y un estudiante puede darse cuenta en medio del test que ha estado respondiendo en la sección equivocada (Timmons et al, 2005). Para niños de primer y segundo grado la experiencia recomienda no utilizar hojas de respuesta, porque la instrucción los puede confundir. Es mejor que respondan directamente sobre la prueba.

IV. Tipos de instrumentos

Existen diversas maneras de clasificar los instrumentos de evaluación de aprendizajes. Por ejemplo, Cueto (2007) propone que una forma de clasificar los sistemas de evaluación es según las implicancias de sus resultados. Así, existen los sistemas de bajas implicancias, que generan información para fines formativos, sin consecuencias directas para los actores involucrados; y los de altas implicancias, que usan los resultados de pruebas para diversos fines que sí tiene consecuencias para los involucrados, como definir incentivos para los docentes, definir la promoción educativa de los estudiantes o informar a la población sobre el rendimiento de los estudiantes en los centros educativos. Estos a menudo son de aplicación censal y están ligados a la rendición de cuentas en educación[13].

Otra manera de clasificar los instrumentos es según sus características técnicas, que es el enfoque que se utilizará en esta sección. Una forma de clasificar bajo este enfoque es según si son medidas directas o indirectas del aprendizaje. Las medidas directas son aquellas en que se observa un producto del trabajo del estudiante, tales como proyectos de investigación o pruebas de conocimiento. Las medidas indirectas, en cambio, son aquellas que no se basan directamente en los trabajos del estudiante sino en las percepciones de estudiantes, docentes, y otros agentes. Este documento aborda principalmente las pruebas de aprendizaje, que son una medida directa del aprendizaje.

A continuación se describen tres clasificaciones muy utilizadas: los instrumentos paramétricos y no paramétricos, referidos a normas o a criterios, y de respuesta abierta versus respuesta cerrada. También se comentan brevemente otros tipos de instrumentos, específicamente las pruebas adaptativas y los de valor agregado. Es conveniente conocer todas estas categorías, que aparecen constantemente en la literatura especializada, y estar familiarizado con las características de cada tipo de instrumentos, para entender sus diferencias y tener claro qué tipo de instrumento sería el más apropiado para el objetivo de evaluación que se tiene.

A Instrumentos paramétricos y no paramétricos

Las pruebas paramétricas están diseñadas para representar a la población general, por ejemplo a la de un determinado país, o de un cierto grupo etario dentro de un país. Normalmente los tests paramétricos están disponibles comercialmente, han sido piloteados y estandarizados en la población completa y ofrecen muchos datos sobre su muestreo, validez y confiabilidad. Los tests

[13] Cueto también plantea que hay grandes críticas a los sistemas de altas implicancias por cuestiones de validez de las inferencias que se pueden realizar. Las objeciones giran alrededor de temas como hasta qué punto son justas las comparaciones entre centros educativos que atienden a poblaciones diferentes, en contextos diferentes, con recursos humanos y materiales diferentes.

no paramétricos, en cambio, han sido diseñados para un grupo específico y no consideran a la población general[14] (Cohen, Manion y Morrison 2000).

Las ventajas de los tests paramétricos son que permiten comparar sub-grupos con la población general, por ejemplo los resultados de una escuela específica contra el puntaje promedio nacional, realizar análisis sofisticados de los datos, y hacer inferencias. Los instrumentos no paramétricos, en cambio, son particularmente útiles para muestras pequeñas y específicas, por ejemplo un cierto curso de una cierta escuela, y pueden ser diseñadas específicamente para lo que se desea medir en esa situación. Además, ya que los análisis que se pueden hacer necesariamente son menos sofisticados, son más fáciles de realizar por cualquier usuario (Cohen, Manion y Morrison 2000).

Un ejemplo de un test paramétrico sería una prueba nacional censal, en que los resultados obtenidos por cada escuela se pueden comparar con los obtenidos por el promedio de las escuelas del país, y de uno no paramétrico, una prueba diseñada por una escuela para diagnosticar los conocimientos en matemática de sus estudiantes al inicio del año escolar.

B Instrumentos referidos a normas y criterios

Evaluación referida a normas

Las evaluaciones referidas a normas son aquellas en que el puntaje obtenido por un individuo se compara con el puntaje obtenido por un grupo, lo que permite definir cómo se ubica el rendimiento de un estudiante o grupo de estudiantes frente a otros (de la misma edad, grado de estudios u otro rasgo en común) (Cueto, 2007; Cohen y Swerdlik, 2009). Entonces, el puntaje se entiende de manera relativa, en comparación a los puntajes obtenidos por otros[15]. Un objetivo importante de este tipo de evaluación es ordenar, o hacer rankings (Cohen y Swerdlik, 2009). Un ejemplo clásico son las pruebas de admisión universitaria, tal como las pruebas SAT.

En este contexto, "norma" se entiende como el comportamiento usual, habitual, esperado, o típico. En el contexto psicométrico, las normas son los datos sobre el desempeño de un grupo

[14] Las pruebas paramétricas hacen dos grandes supuestos sobre la población general y sus características: i) que hay una curva normal de distribución en la población (lo que se observa por ejemplo, en puntajes estandarizados de coeficiente intelectual, o en los logros en lectura de la población en total), y ii) que hay un cero real e intervalos continuos y equivalentes entre los puntajes (a diferencia de los puntajes no ordinales típicos de los cuestionarios, donde no se pueden asumir distancias iguales entre intervalos). Las no paramétricas, en cambio, hacen pocos o ningún supuesto sobre la distribución de la población o sus características, y no asumen una distribución normal, por lo que no permiten comparar con la población general. Por lo tanto, en este caso se debe trabajar con estadísticas no paramétricas (Cohen, Manion y Morrison 2000).

[15] Este modelo proviene de la psicología, donde a menudo las variables de interés, tales como inteligencia, se miden en niveles de intensidad y no por su carencia o dominio absoluto. Aplicado a la educación, este modelo llevó al desarrollo de pruebas que procuraban lograr una distribución normal en el rendimiento y presentaban los datos analizando la posición relativa de un grupo de estudiantes frente a otro. Así, se presentaban resultados indicando en qué percentil se encontraba el rendimiento de un estudiante o grupo de estudiantes y con esto se podía saber qué porcentaje se encontraba por encima y por debajo del grupo descrito (Cueto, 2007).

específico que se usan como referencia para evaluar o interpretar puntajes individuales. Los miembros de esta muestra con que se calcula el puntaje son considerados típicos del grupo, y la distribución de puntajes de este grupo se utiliza como las normas del test contra los cuales se comparan los puntajes individuales (pueden ser puntajes brutos puntajes brutos[16] o estandarizados). El grupo específico (los "otros", con que se comparará el puntaje individual) puede ser tan amplio como la población de un país, o tan específico como los pacientes femeninos de un determinado hospital (Cohen y Swerdlik, 2009).

Existen muchos tipos de normas: por edad, grado, nacionales, locales, de subgrupo, referidas a un grupo de referencia, y de percentil (los datos brutos de la muestra estandarizada convertidos a percentiles) (Cohen y Swerdlik, 2009).

Normar un test a nivel nacional es muy caro, por lo que algunos tests no usan una muestra representativa a nivel nacional para determinar sus normas, sino que solo usan los estadísticos descriptivos de un grupo de individuos en un contexto especifico, a esto se les llama normas de usuarios o programas (Cohen y Swerdlik, 2009).

Evaluación referida a criterios

Tal como se acaba de describir en la evaluación referida a normas, una manera de darle significado al puntaje es compararlo con el de un grupo. La otra forma de darle sentido es comparar el puntaje con respecto a algún criterio o estándar de aprendizaje predeterminado. Por ejemplo, en la evaluación educacional, muchas veces las pruebas se usan para determinar si un individuo ha logrado un cierto nivel de competencia en un campo de conocimiento o habilidad determinada. La evaluación referida a criterios es aquella en que se entrega significado al resultado de un puntaje individual comparándolo con un estándar, que puede ser referido a conocimientos, competencias o habilidades. Aquí el foco está en el desempeño del evaluado, lo que importa es qué puede o no hacer, qué ha aprendido, si cumple o no criterios específicos esperados para su grupo. No importa cómo le ha ido en referencia al resto del grupo, y el puntaje de un individuo no afecta el desempeño de los demás (Cohen y Swerdlik, 2009; Urbina, 2004).

A la evaluación referida a criterios también se le conoce como evaluación referida a contenidos, dominios, objetivos o competencias (Urbina, 2004; Virginia Tech, School of Education, 2014; Cohen y Swerdlik, 2009). Estos distintos nombres se deben a que este tipo de evaluación puede abordar un conjunto de conocimientos en un cierto dominio, demostrados en un test estandarizado, o la demostración de un nivel de competencia, o el cumplimento de ciertos

[16] El puntaje bruto es el primer resultado cuantitativo que se obtiene al corregir un test. Generalmente corresponde al número de respuestas correctas obtenidas. Los puntajes brutos, en sí mismos, carecen de significación y deben ser transformados en puntajes derivados para ser interpretados. Un puntaje bruto sólo puede ser interpretado al ser contrastado con uno o varios grupos normativos, lo que permitirá su transformación a Normas en Puntaje Estándar, Percentiles o Puntaje T.

objetivos de aprendizaje, e incluso puede referirse a la relación entre ciertos puntajes en un test y niveles de desempeño esperados en un cierto criterio (Urbina, 2004).

Este tipo de tests puede entregar datos cuantitativos, por ejemplo porcentajes de respuestas correctas. También puede haber categorizaciones cualitativas, ya sea un todo o nada con respecto a un determinado nivel de logro (por ejemplo, se aprueba o no el examen de conducir), u ofreciendo varios posibles niveles de desempeño intermedio (Urbina, 2004), como sucede en la prueba internacional PISA.

Este tipo de evaluación requiere una cuidadosa descripción de cuál es el estándar esperado y de dónde poner los puntajes de corte (Cohen y Swerdlik, 2009). Es importante tener muy definido el contenido de lo que se desea evaluar y asegurarse de que el test efectivamente evalúa todos los elementos definidos como importante. Para esto, se pueden hacer tablas de especificaciones que detallan la cantidad de ítems por contenido u objetivo de aprendizaje (Urbina, 2004). Estos elementos son muy importantes de considerar si se escoge un test referido a criterios: los criterios y puntos de corte del test se deben relacionar con los objetivos y contenidos de evaluación.

Pros y contras de evaluaciones referidas a normas y criterios

Las ventajas o desventajas de una evaluación referida a criterios o normas tienen que ver con el objetivo de evaluación.

En general la evaluación centrada en los aprendizajes está más relacionada con la evaluación referida a criterios, ya que se liga directamente a objetivos de aprendizaje y a una serie de objetivos de desempeño, en otras palabras, permite evaluar cuánto los estudiantes han logrado los objetivos propuestos, permite medir competencias con respecto a un objetivo instruccional. También la información obtenida puede servirle al mismo estudiante o a los docentes para mejorar en los aspectos identificados como débiles (Virginia Tech, School of Education, 2014). La referida a normas, en cambio, es particularmente útil si lo que se desea es ordenar a estudiantes, escuelas o países, ya sea para rankear o para seleccionar a los mejores. Sin embargo, es difícil establecer de manera absoluta cuánto saben los estudiantes en relación con un determinado estándar o nivel mínimo aceptable (Cueto, 2007).

En la evaluación referida solamente a criterios, una desventaja es que no se tiene información del desempeño relativo del estudiante respecto al resto, y si el test no está diseñado para recoger esta información, se puede perder información sobre los alumnos que están en los extremos, ya sea los más avanzados o los menos avanzados. Por ejemplo, un test podría estar enfocado en medir habilidades básicas de lectura, pero no permitiría identificar a los estudiantes que tienen habilidades lectoras muy avanzadas. En cambio un test referido a normas sí permitiría identificar al estudiante que destaca de entre los demás (Virginia Tech, School of Education, 2014). O al revés, una prueba diseñada para describir aprendizajes avanzados de matemática entrega muy

poca información sobre qué saben hacer los estudiantes que caen bajo el nivel de logro más básico. Según el objetivo de evaluación que se tenga, esto puede ser o no un problema.

Es importante destacar que aunque muchos instrumentos pertenecen a una u otra categoría, ambos enfoques no son necesariamente excluyentes (Cohen y Swerdlik, 2009). Un ejemplo de esto es la prueba PISA, que permite rankear los puntajes y también describe el porcentaje de estudiantes que se ubica en cada nivel de desempeño.

C Instrumentos de respuesta abierta o cerrada

Los instrumentos que se usan para medir aprendizaje pueden separarse en los que son de respuesta abierta y los de respuesta cerrada. Estos últimos implican escoger alternativas dentro de una serie de opciones provistas. El ejemplo más conocido son las pruebas estandarizadas de ítems de selección múltiple. Los instrumentos de respuesta construida, en cambio, son las evaluaciones en que se debe crear la respuesta, tales como ensayos, reportes, u exámenes orales (Center for Assessment and Research, James Madison University, 2014). También caen en esta categoría las pruebas de pregunta abierta, también conocidas como de desarrollo.

Las ventajas de los instrumentos de selección múltiple son que se pueden administrar rápida y fácilmente a grandes números de estudiantes y abordan muchos temas. Sin embargo, ofrecen menos riqueza, menos profundidad y tienden a enfocarse en la capacidad de recordar y otras habilidades cognitivas bajas. Las ventajas de los de respuesta construida son que proveen una información más profunda de los que los estudiantes saben y pueden hacer, pero son caros porque deben ser puntuados por puntuadores entrenados especialmente para ello, lo que además toma bastante tiempo, y en general requieren tiempos de aplicación más largos (Center for Assessment and Research, James Madison University, 2014). Además, si la puntuación no es efectuada correctamente y de la misma manera por todos los correctores, la comparación de los puntajes se ve afectada, lo que no ocurren en el caso de los de respuesta cerrada.

Existen algunas pruebas que combinan ítems de ambos tipos. Por ejemplo, un test puede estar conformado principalmente por ítems de selección múltiple, pero complementar la información recogida con una o dos preguntas abiertas, que requieran que el estudiante redacte un texto y exponga argumentos, lo que permite evaluar habilidades que no se pueden recoger de la misma manera con solo ítems de selección múltiple.

D Otros tipos de instrumentos

Los tests adaptativos

Normalmente los tests están compuestos por un conjunto de ítems que deben ser respondidos por todos los estudiantes. Sin embargo, en algunos contextos puede ser útil utilizar tests adaptativos,

que tienen la particularidad de escoger el grado de dificultad de los ítems que le corresponden a cada estudiante a partir de las respuestas que ya ha dado. Esto logra mucha mayor precisión en la medida y además soluciona el problema de que si el estudiante sabe muy poco, la mayoría de los ítems sea muy difícil para él y tenga que adivinar las respuestas, y que si sabe mucho, la mayoría sean demasiado fáciles. Otras ventajas son que permiten reducir la cantidad de ítems que responde cada estudiante, permiten a los examinados avanzar a su propio ritmo, y los estudiantes sienten que la tarea es desafiante pero realizable.

Los tests adaptativos necesariamente se administran en un computador, y no con lápiz y papel. Al responderse en el computador, no se utilizan hojas de respuesta por lo que el análisis de datos es más fácil (no hay que digitar los datos) y la seguridad de la información es mayor. Son particularmente útiles en evaluaciones a gran escala, en que es esperable que los estudiantes posean una amplia gama de habilidades.

Dentro de las desventajas está que las habilidades requeridas para manejar el computador pueden ser diferentes de las evaluadas por la prueba; el hecho de usar computador puede aumentar o disminuir el nivel de motivación de los examinados; y hay más elementos externos que puede influir en los resultados, como la lentitud del computador, su brillo, etc. También puede haber factores técnicos en el análisis de los datos, relacionados con el cálculo del error estándar[17]. Otras dificultades técnicas son que se requiere un pool de ítems muy grande y que sean de diversos grados de dificultad, lo que puede afectar su calidad, y que es necesario calcular muy bien las reglas de asignación de los ítems a partir de rigurosos pilotajes que informen sobre la dificultad de los ítems (Cohen, Manion y Morrison, 2000).

Pre tests, pos tests y valor agregado

Una pregunta importante en evaluación es no sólo si los estudiantes pueden demostrar aprendizajes al finalizar un cierto curso, año escolar o programa, sino cuánto de esos aprendizajes los obtuvieron durante el curso, año escolar o programa en cuestión.

Esto sugiere la necesidad de medir las habilidades de los estudiantes al ingresar y al egresar del programa, lo que se logra aplicando un pre test y un pos test. Distintos tipos de instrumentos pueden cumplir este propósito: los portafolios son particularmente adecuados. En algunas ocasiones, tests estandarizados o desarrollados para un grupo específico también pueden cumplir este propósito. Sin embargo, si se aplica exactamente el mismo test, la mejoría puede deberse simplemente a que los estudiantes ya lo conocían. Por otra parte, si se administran tests diferentes es difícil asegurar la comparabilidad de ambos (Indiana University Southeast, 2006).

En ocasiones, solo se utilizan pos tests, es decir tests que se aplican una vez finalizado el proceso para identificar si los estudiantes lograron los objetivos de aprendizaje propuestos, sin aplicar un

[17] Para más información sobre cómo calcular este error, ver Thissen, D. (1990) Reliability and measurement precision. In H. Wainer (ed.) *Computer Adaptive Testing: A Primer*. Hillsdale, NJ: Erlbaum, 161–86.

pre test (Virginia Tech, School of Education, 2014). Por ejemplo, si el objetivo es certificar los aprendizajes obtenidos, basta con un pos test, ya que no es parte del objetivo de evaluación identificar cuántos de los aprendizajes se obtuvieron gracias al curso. En otros casos esta pregunta sí es relevante, y para saber cuántos de los aprendizajes se debieron al programa o curso específico sería necesario también aplicarle el pos test a un grupo de control. De lo contrario, no se sabría si los resultados se debieron al programa o curso o a otras variables, tales como el paso del tiempo. Otra alternativa es que algunos proveedores muy especializados proveen pruebas llamadas de "valor agregado", que se administran al principio y al final del año escolar y permiten identificar cuánto de los aprendizajes logrados se debieron al efecto de la escuela o profesor[18].

[18] Para calcular el valor agregado, se deben tener datos agregados, de estudiantes equivalentes, contra los cuales comparar los resultados, para poder estimar cual hubiera sido el desempeño si no se hubiese hecho esa intervención.

V. Consideraciones éticas y justicia de los instrumentos

Esta sección aborda temas relativos a la ética y la justicia en la selección y administración de los instrumentos de evaluación de aprendizajes. Estos son temas fundamentales a nivel valórico, para poder asegurar equidad en las evaluaciones, y a la vez son temas inexorablemente ligados a aspectos técnicos, tales como la validez. Primero se abordan las prácticas éticas de quienes diseñan e implementan instrumentos y de quienes son evaluados. A continuación se discuten aspectos relativos a la justicia y sesgo de las evaluaciones, y se describen los cambios en la situación de evaluación que a veces son implementadas para ajustarse a los estudiantes con necesidades educativas especiales, con la intención de realizar evaluaciones más justas.

A. Consideraciones éticas

Por una parte, es responsabilidad de quien desarrolla la evaluación hacerlo con estrictos estándares éticos en todas sus etapas: en el diseño y selección de la prueba, la revisión de que las normas y estándares sean adecuados para el grupo al que se le administrará, asegurar que las modificaciones para los estudiantes con necesidades educativas especiales han sido hechas correctamente, etc. Ya que muchos de estos elementos requieren conocimientos técnicos, muchas faltas a la ética pueden ser involuntarias, y deberse al desconocimiento técnico. Por otra, también es responsabilidad de quienes son evaluados participar en la evaluación de manera ética, sin embargo hay amplia evidencia de evaluados que intentan falsear los resultados (Cohen, Manion y Morrison, 2000). A continuación se describen las faltas -voluntarias e involuntarias- que pueden ser cometidos por los distintos actores y algunas guías para conducir evaluaciones que cumplan con estándares éticos.

Prácticas éticas por parte de los evaluados

Mientras mayores son las consecuencias de una evaluación, más se presentan conductas poco éticas por partes de quienes están siendo evaluados o serán responsables por los resultados (Cohen, Manion y Morrison, 2000). Se pueden distinguir tres tipos de fraude: a) que el examinado haga trampa al responder, b) el robo de ítems, y c) cuando un tercero cambia los puntajes, ya sea haciéndose pasar por otro para rendir la prueba en su lugar, o cambiando las respuestas dadas por el examinado (Impara y Foster, 2006).

a. La trampa puede manifestarse de distintas maneras: interactuar con otros durante el test a través de celulares u otros medios, usar resúmenes u material prohibido, utilizar más tiempo del permitido, usar calculadoras si están prohibidas, etc. (Impara y Foster, 2006).

b. El robo de ítems consiste en reproducir las preguntas para para luego prestarlas a regalarlas (esto lo pueden hacer los examinados o cualquiera que haya tenido acceso al material). En algunos contextos existen piratas que rinden un test solo para memorizar ítems y venderlos: si trabajan en grupo, logran abordar muchos ítems con solo una aplicación (Cohen y Wollak, 2006; Impara y Foster, 2006). También hay casos de

docentes que entrenan a sus alumnos para que memoricen ciertos ítems y poder así reconstruir una prueba.

c. Con respecto al cambio de las respuestas, existen muchos casos de profesores e incluso de directores que, por presión por tener buenos resultados, cambian las respuestas dadas por los estudiantes en las hojas de respuesta (Cohen y Wollak, 2006; Phillips y Camara, 2006). Otro fenómeno bastante común por parte de las escuelas es pedirle a algunos estudiantes de bajo rendimiento que no asistan el día que hay que rendir el test o directamente entregarle las respuestas correctas durante la evaluación (Cohen y Wollak, 2006).Y aunque es menos común, también hay casos en que estudiantes de cursos mayores rinden un test en lugar del curso que realmente debe rendirlo, es decir, casos grupales de suplantación.

Otra importante fuente de distorsión de los datos relacionados con el rol del docente y director es la preparación que los estudiantes hayan podido recibir para rendir la prueba, también conocido como teaching to the test (Cohen y Wollak, 2006). Esto puede manifestarse de muchas maneras, entre ella: Asegurarse de la cobertura de los contenidos y objetivos del programa que serán incluidos en la evaluación; restringir la cobertura solo a aquellos elementos que serán evaluados; preparar a los estudiantes para rendir ese tipo de examen; practicar con exámenes similares; decirle a los estudiantes con anticipación que entrará en el examen, y practicar con el mismo con el mismo test, sin apoyo del profesor o revisar el test con el profesor (Cohen, Manion y Morrison, 2000).

Ha habido mucho debate sobre qué tan apropiado es el teaching to the test. La barrera entre qué es preparación legítima para un test, y qué no, es difusa. Por ejemplo, en tests con consecuencias para los estudiantes se les deja en desventaja si no se les prepara (Cohen, Manion y Morrison, 2000). De hecho, en pruebas de admisión universitaria parte de las reglas del juego es que los estudiantes se preparan. Sin embargo, la preparación no es algo deseable en la evaluación de programas. Una definición de prácticas poco ética (Gipps, 1994, en Cohen, Manion y Morrison, 2000) es cuando los puntajes aumentan pero las inferencias confiables sobre el desempeño no, y cuando diferentes grupos de alumnos son preparados de manera diferente para el test, dando a algunos ventajas injustas sobre otros. Gipps sugiere que es correcto que los profesores enseñen contenidos más allá de los que serán incluidos en el test, y no preparen para el test: idealmente, solo la mejor instrucción es aceptable. Algunos autores proponen que es inapropiado si hay mucha preparación o ensayo, y es apropiado si se enseñan los contenidos (Phillips y Camara, 2006).

Desde el punto de vista técnico, cuando hay fraude se afecta la validez de los puntajes, ya que los puntajes no reflejan adecuadamente las habilidades medidas por el test. Es decir, son fuente de varianza irrelevante para el constructo siendo medido. Además, el fraude tiene problemas relacionados con la reputación de la evaluación, y con la justicia para quienes no hicieron trampa. En el caso que hay que replicar una prueba, puede ser dañino o agotador para los evaluados tener que ser re evaluados (Impara y Foster, 2006).

Prácticas éticas por parte de los responsables de la evaluación

Los elementos que han sido descritos en esta nota como temas técnicos a tomar en cuenta, junto con otros temas relativos a la calidad que no han sido incluidos aquí, pueden también considerarse deberes éticos (Cronbach 1970; Hanna 1993; Cunningham 1998, todos en Cohen, Manion y Morrison, 2000). Por ejemplo, para que una evaluación sea ética, los instrumentos deben ser válidos y confiables; su administración, puntuación y análisis solo debe ser realizado por personas preparadas y sin intereses creados; se debe controlar el acceso al material; los procesos de puntuación debe ser transparentes.

Al seleccionar instrumentos, es responsabilidad de quien selecciona asegurar que los contenidos están alineados con los objetivos de evaluación y que el formato sea relevante. En el caso de pruebas referidas a normas, la población usada para la muestra a partir de la cual se construyó la norma debe ser representativa a la de la población a la que se quiere generalizar. En el caso de estar evaluando niños pequeños, se debe asegurar que el test sea apropiado a su edad (MCME Code, en Phillips y Camara, 2006).

Muchas veces también está en manos de quien implementa la evaluación prevenir el fraude, poniendo estrictos controles de calidad en el manejo del material antes, durante y después de la aplicación, definiendo que quienes tengan acceso al material no se vean afectados por los resultados y no sean responsables por los resultados que están siendo medidos. Para esto, es esencial seleccionar examinadores apropiados, que no tengan conflictos de interés. Por ejemplo, un error común es usar como examinador al profesor del curso que está siendo evaluado, porque sale barato. También hay que asegurarse que los examinadores seguirán rigurosamente las instrucciones y tienen las capacidades requeridas, por ejemplo, deben poder leer bien en voz alta las instrucciones (McCallin, 2006). Prevenir la distribución de los ítems de los tests previo a su aplicación y desincentivar prácticas poco éticas de preparación para el test son responsabilidad del estado en los casos en que la evaluación es implementada por el estado (Phillips y Camara, 2006).

Otra relevante consideración ética son los derechos de los estudiantes siendo evaluados. Algunos se basan en la legislación nacional o en reglamentos escolares, otros en el sentido común, los estándares éticos comúnmente aceptadas, y el respeto. En algunos casos, se requiere permiso otorgado por los padres o apoderados. Por lo tanto es importante informarse sobre el contexto legal del lugar en que se aplicará la evaluación, junto con considerar el contexto cultural de los estudiantes al seleccionar el método de evaluación, y proveer las adaptaciones necesarias en estudiantes con necesidades educativas especiales (The Joint Committee on Standards for Educational Evaluation, 2003). También es importante respetar su privacidad y dignidad, asegurar que la evaluación no les haga daño, y se debiera solicitar su consentimiento informado antes de participar (Cohen, Manion y Morrison, 2000).

Existen códigos de ética aplicables a los investigadores que hacen evaluación educativa, como los del American Educational Research Association, que incluyen elementos tales como dar información a los participantes de la investigación (Phillips y Camara, 2006). En Estados Unidos existe legislación al respecto, que puede ser considerada como guía para aplicaciones en otras partes. Por ejemplo, El National Research Act de 1974 exige que el estudio haya sido aprobado por una organización externa previo a su implementación, no hacer daño, consentimiento informado, y permiso de los padres o apoderados. El Family Educational Rights and Privacy Act exige privacidad de los registro de los estudiantes, lo que se puede lograr asignándoles un código no ligado a su identidad real. El National Council on Measurement in Education (NCME) ha establecido un código para guiar a sus miembros involucrados en la evaluación en educación y para ofrecer guía a quienes no están afiliados (http://ncme.org). La sección dirigida a quienes seleccionan instrumentos menciona:

- Conducir una revisión exhaustiva de los tipos de evaluaciones y los instrumentos que podrían server para el uso propuestos
- Seleccionar tests basados en la documentación técnica pública sobre su calidad
- Explicitar cualquier relación que puedan tener con los dueños de los tests
- Informar a los tomadores de decisión y posibles usuarios de la adecuación del instrumento para el propósito que se usará, probables consecuencias de su uso, derechos del examinado, costos, requisitos de la evaluación, y limitaciones conocidas.
- Vetar instrumentos que probablemente produzcan resultados inválidos para grupos específicos por motivos de raza, género, nivel socioeconómico, etc.
- Cumplir con todas las precauciones de seguridad en el manejo del material
- Denunciar cualquier intento de presión o manipulación por escoger un cierto instrumento
- Evitar el uso de material de preparación del test que pueda afectar los resultados obtenidos por los estudiantes

La American Educational Research Association, junto con la American Psychological Association y el National Council on Measurement in Education han creado un conjunto de estándares para la evaluación educativa y psicológica, cuya versión revisada de 1999 es ampliamente reconocida., y aborda temas relativos a la construcción de instrumentos, justicia al evaluar, y la administración en distintos contextos (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Sabía usted que…

Desde los primeros exámenes formales en China, hace ya 2000, años se daban casos de hacer trampas en los exámenes, utilizando las mismas técnicas que se ven hoy, tales como hacer resúmenes en los pliegues de la ropa. El castigo podía incluso ser la decapitación.

B. Justicia del instrumento

Otro tema muy relacionado con la ética en la evaluación, es la justicia. ¿Cómo asegurar que un test sea justo? ¿Cómo se relaciona la justicia con el sesgo y con la validez? ¿Cómo se pueden hacer evaluaciones que sean justas para estudiantes con necesidades educativas especiales?

Justicia, sesgo y validez

Desde la década de los '60 la justicia, entendida como el no sesgo de los ítems, ha sido un tema mayor en la sicometría (Zieky, 2006). El sesgo entendido estadísticamente es un error sistemático entre dos medidas cuando estas debieran ser iguales (Camilli, 2006). O, dicho de otro modo, cuando factores irrelevantes o arbitrarios sistemáticamente afectan las inferencias y juicios realizados a partir de una evaluación de una manera que afecta de manera diferente a un estudiante o grupo de estudiantes. Estos factores pueden ser diferencias culturales, de lenguaje, dificultades en el desarrollo o físicas, nivel socioeconómico, género o raza (The Joint Committee on Standards for Educational Evaluation, 2003). Entonces, un test justo busca identificar diferencias, si las hay; y si hay diferencias, estas se deben a las diferencias en la habilidad, no a diferencias en el modo de medir. Es decir, si ambos grupos efectivamente presentan diferencias en su nivel de habilidad, y el test lo detecta, eso no es un test injusto[19] (Camilli, 2006). El sesgo se puede calcular matemáticamente, usando una medida métrica del desempeño de distintos grupos, por ejemplo el DIF (differential item functioning), que permite comparar desempeños en el ítem de dos grupos que tienen el mismo nivel de habilidad[20] (Camilli, 2006; Zieky, 2006).

Aún hoy algunos autores utilizan el término justo como sinónimo de sin sesgo. Entendido así, un test es injusto cuando dos grupos de sujetos obtienen medidas distintas en un instrumento, a pesar de que poseen la misma habilidad en el dominio siendo medido (Camilli, 2006).

Para otros autores, los términos validez, justicia y sesgo de los tests son conceptos diferentes. Mientras que sicométricamente el sesgo es una característica de un test que sistemáticamente impide evaluaciones precisas e imparciales, y se calcula de manera matemática, la justicia es hasta qué punto el texto es usado de manera imparcial, justa y equitativa (Cohen y Swerdlik,

---

[19] Popularmente se entiende que un test justo es aquel que entrega los mismos puntajes para diferentes grupos, por ejemplo de hombres y mujeres. Sin embargo, esta definición está errada. Como se acaba de explicar, para los sicometristas la diferencia entre puntajes de grupos no dice nada sobre qué tan justo es el test: por ejemplo, un grupo de hombres puede medir más que uno de mujeres, pero eso no refleja una injusticia del instrumento que se usó para medir altura, sino que refleja diferencias efectivas entre los grupos. Juzgar la justicia de un ítem por si parece favorecer algunos grupos también es errado, ya que es una percepción subjetiva y no necesariamente correlaciona con el desempeño real de la gente (Zieky, 2006).

[20] Para este análisis es fundamental haber identificado dos grupos que efectivamente tengan el mismo nivel de habilidad, de lo contrario un ítem puede mostrar DIF porque no fue bien hecho el match de habilidad entre ambos grupos, y no porque realmente tenga sesgo

2009). Para otros, los tests "justos" son aquellos libres de sesgo y que además cumplen con estándares reconocidos de administración y ética. Por ejemplo, los tests estandarizados deben ser administrados exactamente de acuerdo a sus instrucciones de aplicación y que todos deben recibir mismas instrucciones (Zieky, 2006; Timmons et al, 2005). Y, ya que la calidad de los puntajes que se obtengan dependerán de la calidad de las normas, se debe averiguar todo lo posible sobre los grupos con los cuales o para los cuales se calcularon las normas. ¿Las normas representan a su muestra de estudiantes? (Bart, 2009).

Zieky (2006) propone que la definición técnica más útil es aquella que liga la justicia con la validez. Cualquier elemento que no es relevante para el constructo, es inválido. Por lo tanto, la justicia requiere que características irrelevantes al constructo de los examinados no afecten los resultados (esto también incluye respuestas emocionales). Es decir, cualquier cosa que reduzca la validez de un ítem reduce su justicia. Ligar la justicia con la validez también implica tomar en cuenta el propósito del test y el uso que se le está dando. Un test justo puede ser usado injustamente. Bajo esta mirada, los fairness reviews debe fijarse en la validez de los ítems, no en si son políticamente correctos. Para poder determinar qué es irrelevante al constructo se hace necesario tener una muy buena definición del constructo a medir (Zieky, 2006).

La justicia también se puede ver afectada si el usuario tiene poca familiaridad con el proceso de evaluación, por ejemplo, una persona que nunca ha usado una hoja de respuesta y no entiende su funcionamiento puede ver afectado su puntaje. Otros factores pueden ser la familiaridad con el idioma del test, con su modo de administración (lápiz y papel o computador) y experiencias previas o de su familia con la evaluación. Realizar sesiones de práctica puede ser útil en estos casos. Los administradores de los tests también debieran a estar pendiente de si algún niño, especialmente los más pequeños, presentan dificultades con las hojas de respuesta (Timmons et al, 2005).

Para que un test sea usado de manera justa un elemento esencial es considerar las diferencias culturales y lingüísticas de los evaluados. En Estados Unidos, ha habido un énfasis en asegurar justicia para distintos grupos: razas, minorías, género, discapacidad, religión, nivel socioeconómico y edad. No se puede dar por hecho que a los miembros de diversas comunidades les parecerá que ciertos elementos o ítems son adecuados; ni asumir que una misma prueba sirve para todos, cuando se está evaluando a personas de distintos backgrounds culturales y lingüísticos, ni que porque un test ha sido traducido a otro idioma, es exactamente igual al original en todos los otros aspectos, ni que los supuestos a la base de un test afectarán a los grupos de distintas culturas de la misma manera (Cohen y Swerdlik, 2009). El fairness review guidelines (www.ets.org) propone como lineamientos tratar a la gente con respeto; minimizar el efecto de habilidades irrelevantes al constructo; evitar material ofensivo o innecesariamente controversial; usar terminología apropiada para referirse a la gente; evitar estereotipos, y representar diversidad de personas en los ejemplos. Estas guías sirven para cualquier cultura, pero deben interpretarse de acuerdo a cada cultura. Por ejemplo, algo puede ser inapropiado en Arabia Saudita y aceptable en Suecia (Zieky, 2006). Siempre es una buena idea analizar los

ítems disponibles para ver si para los estudiantes específicos puede contener material ofensivo o las palabras puedan tener más de un significado (Bart, 2009).

Sabía usted que…

En los países de LAC se usan diferentes medidas para la comida, por ejemplo comprar manzanas por kilo o por unidades, y arroz por caja o por kilo, lo que puede afectar la comprensión de pruebas de matemática basadas en ejemplos de la vida diaria.

Acomodaciones

Un elemento muy relacionado con la justicia de las evaluaciones son los cambios a las condiciones de aplicación para permitir que estudiantes con necesidades educativas especiales (NEE) puedan participar de las mismas evaluaciones de sus compañeros y sean evaluados de manera justa.

Estos cambios, ya sean acomodaciones o modificaciones, han sido controversiales porque es difícil hacerlo sin interferir con el constructo siendo evaluado (Zieky, 2006; Thurlow, Thompson, y Lazarius, 2006; Phillips y Camara, 2006).

Tanto las acomodaciones como modificaciones implican cambios a la situación estandarizada de administración. Las acomodaciones no interfieren con el constructo siendo evaluado, y pueden referirse a la presentación (por ejemplo, braile en vez de palabra escrita); tiempos de respuesta para estudiantes con dificultades con el lenguaje; ayuda para marcar las hojas de respuesta en el caso de niños que no puedan marcarlas por sí mismos; o cambios a la locación (hacerlo en un primer piso para que pueda acceder un estudiante con movilidad reducida). Para que sean correctas, estas acomodaciones deben ser necesarias para la persona con NEE para poder rendir la prueba y no se debe afectar la validez y comparabilidad de sus puntajes. Estas acomodaciones no deben relacionarse con habilidades relativas al constructo siendo medido. Por ejemplo, ayudar a un parapléjico a marcar las respuestas en una prueba de matemática es solo una acomodación (Phillips y Camara, 2006). Sin embargo, existe poca evidencia investigativa sobre los efectos de las acomodaciones. Es difícil saber si lo que se elimina es irrelevante al constructo ya que no se sabe qué afecta o no el constructo. Por ejemplo, todas las pruebas tienen un componente de comprensión lectora (Thurlow, Thomposn, Lazarius, 2006). Las modificaciones, en cambio, sí cambian el constructo siendo evaluado, por lo tanto se pierde la estandarización y la comparabilidad.

Este es un tema que no sólo tiene aristas técnicas, sino también políticas y legales. Muchas veces existe presión por grupos de opinión para que las evaluaciones cuenten con acomodaciones apropiadas para no excluir ni perjudicar a estudiantes con NEE.

Además, en Estados Unidos existe mucha legislación sobre este tema, que es necesario tomar en cuenta si se administran evaluaciones en ese país, y que si se está en otro país, es recomendable revisar si existe legislación nacional al respecto (Thurlow, Thompson & Lazarius, 2006).

VI Crear un instrumento o utilizar uno ya existente

Existen muchos instrumentos disponibles para su uso comercial que pueden ser usados para propósitos evaluativos. Desde el punto de vista práctico, utilizar un test que ya existe permite ahorrar mucho tiempo y recursos (Center for Assessment and Research, James Madison University, 2014; Cohen, Manion y Morrison, 2000). Otras ventajas son que en general son instrumentos técnicamente sólidos, es decir, han sido piloteados y estandarizados en una población detalladamente descrita, declaran su validez y confiablidad, cubren una amplia gama de contenidos, tienden a ser tests paramétricos por lo que se pueden hacer análisis sofisticados, incluyen instrucciones detalladas para su administración, en general son fáciles de administrar y puntuar, y en general incluyen orientaciones para la interpretación de los resultados (Cohen, Manion y Morrison, 2000).

Dentro de las posibles desventajas están: son caros, muchas veces están dirigidos a una población muy específica y pueden no adecuarse al propósito evaluativo requerido; algunos tienen una disponibilidad restringida por lo que puede ser necesario afiliarse a cierta institución para usarlo, lo que puede exigir cumplir con ciertos requisitos; y los tests disponibles por definición están pensados para una población general y no hechos a medida para necesidades locales.

Para utilizar uno de estos tests se debe estar seguro que los objetivos, propósitos y contenidos de dicho test están alineados con los objetivos de evaluación. Los Standards for Educational and Psychological Testing declaran que para que un investigador decida si le conviene usar un instrumento ya existente, la regla de oro es que debe poder demostrar adecuación al propósito (Cohen, Manion y Morrison, 2000). Sin embargo, es difícil encontrar un instrumento que se ajuste exactamente a los objetivos específicos de un programa (Center for Assessment and Research, James Madison University, 2014).

Algunos errores comunes que se cometen al seleccionar instrumentos, que afectan su validez y por lo tanto deben evitarse, son: usar determinado instrumento porque tiene buena fama o ha sido usado antes; usar información porque está disponible; usar métodos con los que no se está familiarizado, sin capacitarse adecuadamente; no proveer de adecuaciones a estudiantes con NEE o que no hablan el idioma (Joint Committee on Standards for Educational and Psychological Testing, 1999).

Si se decide diseñar un instrumento, la principal ventaja es que se crea a medida para estar perfectamente alineado con los objetivos del programa (Center for Assessment and Research, James Madison University, 2014) y se adecuará con precisión al contexto local e institucional (Cohen, Manion y Morrison, 2000). Otro factor a considerar es que aunque diseñar un instrumento implica mucha inversión de recursos, el instrumento le pertenece a quien lo construye, por lo que puede ser una buena inversión si se quiere aplicar a muchos estudiantes o por un periodo largo de tiempo, mientras que si se usa uno ya existente muchas veces hay que

pagar por cada aplicación (Center for Assessment and Research, James Madison University, 2014).

Sin embargo, elaborar instrumentos es caro, lento, y porque probablemente será no paramétrico, el rango de posibles análisis será más limitado que en el caso de un test paramétrico (Cohen, Manion y Morrison, 2000). Además, muchas veces no se cuenta con personal especializado que pueda diseñar un instrumento técnicamente riguroso (Center for Assessment and Research, James Madison University, 2014).

En Estados Unidos existen una serie de organizaciones que han desarrollado estándares o recomendaciones con prácticas para la elaboración de tests, su interpretación y uso. Tal vez los más famosos son los estándares para los tests psicológicos y educativos, esfuerzo conjunto de The American Psychological Association y el National Council on Measurement in Education, siendo su última versión la del 1999 (Buckendahl & Plake, 2006). Un capítulo de los Estándares para la Evaluación Educativa y Psicológica (Joint Committee on Standards for Educational and Psychological Testing, 1999) se refiere a la información que deben dar los elaboradores de los tests, que en general es tomado como guía por los tests disponibles comercialmente. El objetivo de estas recomendaciones es proveer a los usuarios de la información relevantante (Buckendal y Plake, 2006). Destaca la necesidad de declarar:

a. El propósito del test, los usos sugeridos de sus resultados, incluyendo el grupo etario y cualificaciones de quienes interpretan los datos. Este propósito se debe comparar con el de la evaluación (Buckendal y Plake, 2006).
b. información sobre cómo se construyó el test
c. información técnica sobre normas, escalamiento, información detallada sobre la muestra con la que se construyó la norma (hay que compararlo con el grupo). Evidencia sobre la generalizabilidad de los puntajes y validez (Buckendal y Plake, 2006).

VII Conclusiones

Tal como se anticipó en la introducción, escoger el instrumento de evaluación más adecuado para el propósito evaluativo que se tenga no es una tarea fácil. Convergen una serie de consideraciones técnicas, éticas, prácticas, y en ocasiones políticas, que además pueden señalar en diferentes direcciones. Probablemente se tendrán que hacer compromisos, y no existirá un instrumento ideal, pero es esencial tomar una decisión informada y tener claros los riesgos que se están corriendo.

La principal consideración a tener en cuenta es que el instrumento escogido debe ser coherente con el objetivo de evaluación que se tiene. Esto puede parecer evidente, pero en muchos casos las personas se ven tentadas a escoger un instrumento por otros motivos, tales como porque está a mano, lo han usado antes, lo saben usar, o un grupo de referencia importante lo utiliza. También existen presiones políticas para utilizar ciertos instrumentos o tipos de instrumentos.

Una gran dificultad es que no hay una solución única aplicable a todas las situaciones, ni un checklist de elementos que se puedan revisar rápidamente cada vez que se revisa un instrumento. Cada objetivo evaluativo requiere que se haga un análisis detallado para ver si el instrumento es adecuado para ese contexto particular, considerando si el instrumento efectivamente evalúa lo que el objetivo de evaluación busca, si la población para la que fue diseñada el instrumento es equivalente a la población a la que se desea evaluar, y también tomando en cuenta los recursos que hay disponibles. Incluso la calidad de los instrumentos no es algo independiente del contexto, ya que las definiciones más recientes de validez la no la entienden como la validez del instrumento ni de sus puntajes, sino la de la interpretación de los resultados del test para determinados usos, en determinados contextos, y además la validez y confiabilidad se juegan también en la aplicación de los instrumentos, no solo en su diseño.

Es especialmente relevante poder aplicar los criterios éticos relacionados con que el test escogido sea lo más justo posible. Es muy difícil separar estos elementos de los elementos técnicos, ya que muchas veces lo que es correcto desde el punto de vista de la validez y la sicometría, coincide con lo que es ético y justo. Por lo tanto, es importante contar con los conocimientos técnicos para poder tomar decisiones informadas: de lo contrario, se cae en el riesgo no solo de tomar decisiones técnicamente incorrectos, sino también éticamente cuestionables, en que no se está entregando a todos los grupos de estudiantes las mismas posibilidades de demostrar sus conocimientos, o se permite a algunos grupos obtener puntajes inflados gracias a haber cometido fraude. Qué es los más justo posible también depende de cada situación particular.

En el caso que se decida que los instrumentos disponibles no cumplen con los requisitos que permitan justificar su uso, se hará necesario diseñar un instrumento especialmente. Esto puede ser una gran oportunidad para desarrollar un instrumento que se ajuste perfectamente a las necesidades del objetivo evaluativo, y de aplicar las diversas recomendaciones planteadas en este documento. Por otro lado, también requerirá una inversión importante y contar con personas especializadas que lo puedan desarrollar.

Bibliografía

Banco Interamericano de Desarrollo, División de Educación, 2013. *Documento de Marco Sectorial de Educación y Desarrollo Infantil Temprano.*

Bart, M. 2009. What You Need to Know When Evaluating Assessment Instruments. Available at http://www.facultyfocus.com/articles/educational-assessment/what-you-need-to-know-when-evaluating-assessment-instruments/

Brennan, R. 2006. Perspectives on the Evolution and Future of Educational Measurement. In *Educational Measurement* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

Buckendahl C., & Plake, B., 2006. Evaluating tests. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Camilli, G., 2006. Test Fairness. En *Educational Measurement,* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

Center for Assessment and Research, James Madison University. 2014. The Programme Assessment Support Services. Downloaded September 10th from http://www.jmu.edu/assessment/pass/assmntresources/instruments.htm#ExistingInstruments

Cueto, S. 2007. Las evaluaciones nacionales e internacionales de rendimiento escolar en el Perú: balance y perspectivas. En *Investigación, políticas y desarrollo en el Perú. Lima: GRADE*. p. 405-455. Available at http://www.grade.org.pe/download/pubs/InvPolitDesarr-10.pdf

Cohen, A. & Wollak, J. 2006. Test Administration, Security, Scoring, and Reporting. In *Test Administration, Scoring and Reporting.* Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger publishers, Westport.

Cohen, L., Manion, L., & Morrison, K. 2000. *Research Methods in Education* (6th edition). London, RoutledgeFalmer.

Cohen, R. & Swerdlik, M. 2009. *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th Edition). Boston: McGraw-Hill Higher Education

Darr, C., 2005. A Hitchhiker's Guide to Validity. Available at: http://toolselector.tki.org.nz/Assessment-fundamentals/Criteria-for-choosing-an-assessment-tool

Haertel, E. 2006. Reliability. In *Educational Measurement,* (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

Joint Committee on Standards for Educational and Psychological Testing, 1999. *Standards for Educational and Psychological Testing.* American Educational Research Association, Washington DC.

Indiana University Southeast. 2006. The Indiana University Southeast Faculty Assessment Handbook. Available at: http://www.ius.edu/oie/program-assessment/assessment-handbook.html

Institute for Digital Research and Education, UCLA (n.d.). SPSS FAQ. What does Cronbach's alpha mean? Available at: http://www.ats.ucla.edu/stat/spss/faq/alpha.html

Impara, J. & Foster, D., 2006. Item and Test Development Strategies to Minimize Fraud. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

McCallin, R., (2006). Test Administration. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Ministry of Education of New Zealand, 2014. *Criteria for choosing an assessment tool.* Downloaded July 20th, 2014 from http://toolselector.tki.org.nz/Assessment-fundamentals/Criteria-for-choosing-an-assessment-tool

National Council on Measurement in Education (NCME), 2104. Code of professional responsabilties in educational measurement. Available at: http://ncme.org/resource-center/code-of-professional-responsibilities-in-educational-measurement/

Ontario Ministry of Training, Colleges and Universities, 2011. Selected assessment Tools. Downloaded July 20th, 2014 from

http://www.tcu.gov.on.ca/eng/eopg/publications/OALCF_Selected_Assessment_Tools_Mar_11.pdf

Phillips, S., & Camara, W., 2006. Legal and Ethical Issues. In *Educational Measurement*, (4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on education. Praeger publishers, Westport.

The Joint Committee on Standards for Educational Evaluation, 2003. *The Student Evaluation Standards.* Corwin Press Inc. Thousand Oaks, California

Timmons, J., Podmostko, M., Bremer, C., Lavin, D., & Wills, J. (2005). *Career planning begins with assessment: A guide for professionals serving youth with educational & career development challenges (Rev. Ed.).* Washington, D.C. Downloaded from http://www.ncwd-youth.info/career-planning-begins-with-assessment

Thurlow, M., Thompson, S., & Lazarius, S., 2006. Considerations for the administration of tests to special needs students: accommodations, modifications, and more. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Urbina, S. 2004. *Essentials of Psychological Testing*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Virginia Tech, School of Education, 2014. *Introduction to Instructional Design. Lesson 7, Assessment Instruments*. Available at: http://www.itma.vt.edu/modules/spring03/instrdes/lesson7.htm

Webb, N., Shavelson R., & and Haertel, E., 2006. Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics, Volume 26: Psychometrics*. Rao, C & Sinhara, R. eds. Available at: http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20Hdbk%20of%20Statistics.pdf

Wilson, M. 2005. *Constructing measures, an item response modeling approach*. Lawrence Erlbaum Associates Inc., Publishers. Mahwah, New Jersey.

Yen, W. & Fitzpatrick, 2006. Item Response Theory. In *Educational Measurement*,(4th edition). 2006. Brennan R., Ed. Sponsored jointly by the National Council on measurement in Education and American Council on Education. Praeger Publishers, Westport.

Zieky, M., 2006. Fairness reviews in assessment. In *Handbook of Test Development*. Downing, S., & Haladyna, T., Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.