EVALUACIÓN Y POLÍTICAS EDUCATIVAS EN GUATEMALA: PROMOVIENDO UNA POSTURA COMÚN

Assessment and educational policies in Guatemala: Fostering a common approach

Álvaro M. Fortin Morales*
Ype H. Poortinga**
Fons J. R. van de Vijver***

Resumen

Los encargados de políticas públicas buscan, cada vez con más frecuencia, utilizar los resultados de las evaluaciones educativas como fuente de información para retroalimentar y sustentar las decisiones que toman. En respuesta, las unidades de evaluación deben emplear métodos y procedimientos que les permitan proveer el tipo de información y conclusiones que contribuyan a tomar decisiones de política educativa. Estos métodos también tendrán que asegurar la validez de las evaluaciones. La evaluación educativa presenta un gran reto para países heterogéneos como Guatemala, que además experimentan con políticas dedicadas a la equidad. Es necesario que la evaluación se encuentre libre de sesgos causados por factores indeseados y que refleje con exactitud las áreas curriculares evaluadas. Las perspectivas actuales sobre validez sugieren contar con un plan de interpretación de resultados. Los actores clave locales (en especial padres, madres y docentes) y los administradores de política pública interpretarán los resultados de la evaluación de acuerdo con sus experiencias y formación. Sugerimos en este artículo que es necesario alinear las posturas que guían la política educativa con las acciones locales y con la validación de los instrumentos. También proponemos que la información que fundamenta esta perspectiva común debe ser diseminada en un formato accesible a todos los interesados.

Palabras clave: Guatemala, evaluación educativa, políticas educativas, validez, actor local

Abstract

Educational assessment is on the increase as policy makers seek feedback and support for the decisions they make. They turn to assessment as a source of such information. Assessment units then need to adopt methods and procedures enabling them to provide the type of information and conclusions that will enhance policy decisions. Furthermore, these methods will have to assure the validity of assessments. In culturally heterogeneous countries with equity-seeking policies, such as Guatemala, assessment poses several challenges. Validity in diverse societies requires that tests be free of bias from any unwanted factors and accurately reflect the school subjects that are being tested. Current perspectives on validity suggest that there should be well-supported plans for interpreting the data produced by the assessment. Local stakeholders (especially parents and teachers) and policy makers will interpret assessment results according to their background and experience. We suggest that the alignment of perspectives that guide policy, local actions as well as validation of the assessment is required, and that information on this unified perspective be disseminated to all in an accessible form.

Key words: *Guatemala, educational assessment, educational policy, validity, stakeholder*

* Universidad del Valle de Guatemala, Guatemala, Universidad de Tilburg, Países Bajos, alvarofortin@gmail.com
** Universidad de Tilburg, Países Bajos.
*** Universidad de Tilburg, Países Bajos.

Ha incrementado la aceptación de la evaluación estandarizada como herramienta para dar seguimiento a la calidad educativa. En Guatemala, en donde ésta es relativamente reciente y el Estado batalla contra condiciones de marcada desigualdad y pobre calidad educativa, su aplicación presenta un gran reto. Por ello, las evaluaciones deben acompañarse de información que describa sus alcances, aplicaciones y limitaciones en un estilo acorde a la audiencia a la que se dirige. Para que esta comunicación tenga un impacto en las estrategias de fomento a la calidad educativa es necesario contar con una postura común expresada en un lenguaje coherente y de amplio acceso. Dicha perspectiva debe ser comprendida (e idealmente compartida) por los encargados de las políticas, los actores clave locales (padres y madres como representantes de los intereses del estudiante y los docentes), los administradores y los diseñadores de pruebas. Esto asegurará que las políticas, las acciones y los sistemas de rendición de cuentas se encuentren en una misma vía, permitiendo que la evaluación afecte la toma de decisiones. Esto requiere que todos los involucrados negocien una plataforma común de acción. Este artículo presenta la experiencia guatemalteca y se sugiere que dicha plataforma debe basarse en los mismos argumentos que justifican la relación entre la política pública y los dominios que intenta modificar, para ser útil como herramienta de mejora educativa en contextos heterogéneos.

El artículo presenta primero información general de Guatemala, la historia de la evaluación en el país y su evolución en respuesta a las políticas educativas. A esto le sigue una discusión sobre el impacto que los objetivos de política tienen sobre la evaluación y la validez. El artículo cierra con una discusión sobre la necesidad de diseminar la información y de promover una plataforma común entre interesados.

Guatemala: antecedentes de país y contexto político

Desde que la primera capital guatemalteca se fundó en 1524, la sociedad se encontraba estratificada y la población maya ocupaba los escaños inferiores. Diversas condiciones sociales, económicas y políticas reprodujeron dicha situación hasta la actualidad, aun cuando la base legal para la misma ya hubiese desaparecido. Gradualmente incrementaron los conflictos sociales y la demanda de equidad social hasta que la guerra civil explotó en 1960, perjudicando principalmente al área rural. En diciembre de 1996 se firmaron Acuerdos de Paz que exigen tomar acciones para contrarrestar la desigualdad (Gobierno de Guatemala, Unidad Revolucionaria Nacional Guatemalteca, Naciones Unidas, 1996). También se publicó un documento de propuesta para la Reforma Educativa, el cual presenta lineamientos generales que buscan eficiencia, eficacia y equidad (CCRE, 1996). Guatemala también participa de los acuerdos de educación para todos convenidos en Dakar en el año 2000.

Guatemala es un país étnicamente diverso. Los grupos oficialmente reconocidos como "pueblos" son el ladino o mestizo, maya, garinagu[1] y xinca. El idioma oficial es español, pero otros 23 idiomas también se encuentran en uso (21 idiomas mayas, xinca y garífuna) (Ethnologue, 2005; Richards, 2003). Los grupos indígenas suman cerca de la mitad de la población (Beckett & Pebley, 2002; Jiménez Sánchez, 1998; World Factbook, 2007). La población maya se localiza principalmente en áreas rurales, las cuales son más pobres que las urbanas. El índice de distribución de riqueza denominado Gini es 55,1 para el país (UNDP, 2006), pero mientras que desciende a 40,5 para áreas urbanas, llega a 63,0 para áreas rurales (Porta & Somerville, 2006). La tasa de acceso a la educación es superior en áreas urbanas y para comunidades ladinas (Álvarez & Schiefelbein, 2007). La población maya tiende a pertenecer a un nivel socioeconómico inferior al promedio (Beckett & Pebley, 2002; Esquivel Villegas, 2006). Los ingresos de estudiantes que se han graduado del nivel secundario son más altos para ladinos de género masculino que para otros grupos (Porta & Laguna, 2007). Las mujeres mayas son el grupo con el menor acceso a educación (Esquivel Villegas, 2006). Las diferencias de ingreso asociadas a la etnia incrementaron entre 1988 y 1995, a pesar del crecimiento económico del país en esa época (Beckett & Pebley, 2002). Como consecuencia, las políticas orientadas a la equidad se han concentrado en las mujeres indígenas que habitan áreas rurales.

Los objetivos de la política educativa

La política educativa guatemalteca ha experimentado con nuevas estrategias para responder a las demandas de equidad en la educación. La educación bilingüe intercultural y la reforma curricular son acciones que se han ejecutado, que han traspasado gestiones y aún se encuentran vigentes.

Antes de 1960 el modelo dominante para atender a poblaciones bilingües era el de "castellanización". Éste desalienta el uso de los idiomas indígenas y favorece el uso del español, bajo el supuesto de que esto facilitará la integración de los individuos y la participación de la nación en el "mundo moderno" (Antillón Milla, 1997). Hacia la década de 1980, con la creciente presión por implementar programas en favor de la equidad, se cambió esa perspectiva por el modelo de Educación Bilingüe Intercultural (EBI). Se espera que en aulas EBI los estudiantes sean educados en la lengua materna durante los primeros tres años de escolaridad para asegurar que sus procesos cognoscitivos y lingüísticos básicos estén bien establecidos antes de iniciar procesos educativos en español. La EBI mostró resultados iniciales positivos (Chesterfield, Rubio & Vásquez, 2003).

[1] También conocido por su variante en singular: garífuna.

La reforma curricular tomó más tiempo. En el 2005 el Ministerio de Educación publicó un nuevo currículo basado en las competencias que se espera que los estudiantes logren, reemplazando la tradicional organización por contenidos (Ministerio de Educación de Guatemala [Mineduc], 2005). Los estándares educativos se publicaron en el año 2006 para clarificar el desempeño esperado de los estudiantes (Ministerio de Educación de Guatemala [Mineduc] & Programa Estándares e Investigación Educativa-USAID, 2006). La evaluación basada en estándares requiere alinear los ítems de las evaluaciones a los contenidos curriculares y a los niveles de desempeño esperados. En el caso de poblaciones heterogéneas se requieren condiciones adicionales, dado que los instrumentos deben permitir comparaciones válidas entre grupos y con respecto a estándares equivalentes para todos los grupos.

breve historia de las evaluaciones nacionales en guatemala

La política educativa ha influido de varias maneras la forma en que se realiza la evaluación educativa estandarizada en Guatemala. En 1997 y 1998 se evaluaron muestras representativas de estudiantes de tercer y sexto grados (de Baessa, 1997). En 1999 y 2000 también se desarrollaron instrumentos en cuatro idiomas maya (k'iche', kaqchikel, q'eqchi' y mam) para evaluar a estudiantes de tercer grado de escuelas EBI (de Baessa, 1999a, b, 2000a, b). En el 2001 la muestra tuvo que ser reducida por razones de presupuesto. Dada la prioridad otorgada a las políticas orientadas a la equidad, sólo se evaluaron áreas rurales (de Baessa, 2001a, b).

Otros actores también han ejercido influencia sobre las tendencias en evaluación. En Guatemala cerca del 15% de las instituciones educativas del nivel primario son privadas, pero las muestras del período comprendido entre 1997 y 2000 no las incluyó. Cuando en 1999 se intentó cubrirlas, una asociación de las mismas instituciones antepuso un amparo legal contra las evaluaciones y con ello logró detenerlas. La evaluación en instituciones privadas sólo se intentó nuevamente en el año 2006, cuando el país participó en un estudio internacional que requería que este sector se incluyera en el muestreo. Hubo menos presión ante esta evaluación y se logró conducirla de manera satisfactoria.

Las áreas curriculares evaluadas han sido matemática y lectura, como representantes de las habilidades más básicas. Se elaboraron versiones diferentes de las pruebas para las áreas urbana y rural. Dado que antes del 2005 el Ministerio de Educación no había publicado un currículo detallado, el contenido de las pruebas lo seleccionaron docentes con base a su juicio de lo que era apropiado para el grado de interés. Las pruebas fueron desarrolladas para comparar el desempeño de los estudiantes con el desempeño de su

cohorte, pero no para compararlos contra un criterio externo (i.e., las evaluaciones eran referidas a la norma).

Durante el período 1997 a 2000 se utilizaron siempre procedimientos de análisis similares, tanto para la muestra total como para submuestras, con el objeto de estudiar mejor a los grupos vulnerables. Los resultados fueron consistentes entre años (ver de Baessa, 1997, 1998, 1999a, 1999b, 2000a, 2000b, 2001a, 2001b). Los niños obtuvieron mejores resultados que las niñas, tanto en lectura como en matemática. Los estudiantes con sobreedad (estudiantes que han repetido grado) tienden a obtener resultados por debajo del promedio nacional. Los resultados del área urbana muestran una desviación estándar por arriba de las áreas rurales. Las poblaciones mayas mostraron resultados por debajo de los obtenidos por los estudiantes ladinos, pero en las evaluaciones en español los resultados en matemática mostraron mayor semejanza a la tendencia nacional que los resultados en lectura.

Debido a limitaciones financieras no hubo evaluación en los años 2002 y 2003. El préstamo del Banco Mundial que financió la fundación del programa de evaluación había caducado. Sin embargo, la administración 2004-2008 decidió establecer un sistema de aseguramiento de la calidad que requería que los estudiantes fueran evaluados. Con ocasión de la nueva adopción de la evaluación se hizo una revisión de instrumentos y procedimientos en preparación a la publicación del nuevo currículo. Esta fue la primera ocasión en la que los estudiantes inscritos en el primer grado fueron evaluados con un instrumento que compara su rendimiento con el de un estándar de lo que se considera aceptable (i.e., evaluaciones referidas al criterio) (Crocker & Algina, 1986).

El currículo para los niveles de primaria se publicó en el año 2005. La nueva tarea era alinear las pruebas con este currículo para que los resultados informaran acerca de los logros en las competencias establecidas por el mismo. Para esto las pruebas deben tener un grupo suficientemente grande de ítems para representar con exactitud el área curricular bajo evaluación. Se elaboraron diferentes formas de pruebas que contienen ítems comunes para poder realizar comparaciones entre ellas. El número de ítems relacionados con cada dominio se tornó muy grande para que un solo estudiante los pudiera responder todos. Por tanto, se realizó un procedimiento de administración en "espiral". Esto se refiere a la distribución de más de una forma de la evaluación entre estudiantes, para asegurar que todos los ítems de un dominio sean respondidos, al tiempo que se evita que un solo estudiante tenga que responder a instrumentos de demasiada extensión (Dings, Childs & Kingston, 2002). Al procedimiento se le denomina de "espiral" debido a que las versiones se entregan en orden consecutivo a estudiantes que se sientan continuo o detrás del estudiante previo, iniciando nuevamente con la primera forma en cuanto se ha repartido la última, creando así un espiral en las filas de los estudiantes.

Bajo estas condiciones el análisis psicométrico clásico que se utilizó en años anteriores es inadecuado. El análisis clásico se basa en la frecuencia relativa de respuestas correctas en un ítem y la puntuación de varios ítems en una combinación lineal. Las puntuaciones totales se estiman sumando el resultado en ítems individuales (Nunnally & Bernstein, 1995). Para este tipo de análisis es adecuado trabajar con medias y puntuaciones de consistencia interna. Esto es insuficiente cuando grupos de estudiantes diferentes responden pruebas con diferentes ítems, porque esto no permite las comparaciones entre versiones y grupos. Dado que dichas comparaciones son importantes para modelar programas educativos, fue necesario encontrar una metodología alterna.

La Teoría de Respuesta al Ítem (TRI) fue introducida para responder al problema antes planteado. La TRI estima el patrón de respuesta de estudiantes de diferentes niveles de habilidad, aun cuando los estudiantes no responden al mismo grupo de ítems. El supuesto es que una variable latente común es subyacente a todos los ítems. Dicha suposición provee un marco de referencia para comparar el desempeño de los evaluados que han tomado pruebas diferentes o que se encuentran a niveles diferentes de desempeño (Crocer & Algina, 1986). En Guatemala se utilizó TRI para aparear las formas utilizadas en el año 2006.

También se buscó un método para detectar el nivel del logro de los estudiantes con relación a los logros planteados por el currículo. Los estándares educativos fueron publicados en el año 2006 (Ministerio de Educación de Guatemala & Programa Estándares e Investigación Educativa –USAID–, 2006). La primera fase incluyó la definición de condiciones para la redacción de ítems y el ensamblaje de las pruebas. Después se asignaron niveles de desempeño con base al procedimiento "Bookmark"[2], que se basa en consultas de expertos. En este caso los expertos eran docentes en servicio, a quienes se les solicitó que indicaran el nivel de desempeño aceptable en ítems específicos (MacCann & Gordon, 2004; Buckendahl, Smith, Impara & Plake, 2000; Kiplinger, 1997). Este procedimiento permitió establecer el punto de corte para identificar a los estudiantes que alcanzaron el estándar.

Los hallazgos para 2006, analizados con este método, fueron consistentes con reportes previos. Las áreas rurales, estudiantes femeninos y grupos indígenas obtuvieron los desempeños más desfavorables, tanto en lectura como en matemática (Moreno Grajeda, Gálvez-Sobral, Bedregal & Roldán, 2008). En el año 2006 los datos también fueron utilizados en un esfuerzo sistemático por comunicar y utilizar los resultados al nivel local. Cada escuela evaluada recibió un manual y una tabla de los resultados de

[2] Se le denomina Bookmark como metáfora asociada a un separador o marcador de libros, dado que los docentes deben marcar y separar en un listado de ítems colocados del más fácil al más difícil, el punto de corte que divide el desempeño aceptable del desempeño que está por debajo del estándar.

la escuela y su departamento[3]. Las escuelas que no participaron en la evaluación sólo recibieron la tabla de resultados del departamento. El manual ofrecía sugerencias para desarrollar un plan educativo institucional basado en los resultados de la evaluación y el juicio local sobre la situación de la escuela.

Esta breve historia de la evaluación guatemalteca ilustra cómo los procesos de evaluación intentaron responder a las políticas emergentes. La unidad de evaluación elaboró pruebas en idiomas mayas para responder a una política que busca la equidad impulsando la educación bilingüe. Los antecedentes históricos también muestran cómo diversos actores clave pueden intervenir. Por varios años las muestras no fueron de cobertura nacional, aun cuando geográficamente incorporaban a todo el país, debido a la falta de cooperación de las instituciones privadas. Se intentó incrementar el uso y aceptación de la información comunicando los resultados a actores clave locales, pero aún es necesario examinar el grado de impacto que esta diseminación tuvo en la calidad educativa.

La interacción entre evaluación y política educativa

Se discutieron ya varios cambios metodológicos que fueron realizados como respuesta a las tendencias en la política educativa. Por ejemplo, la inclusión de grupos excluidos requirió que el alcance de las poblaciones evaluadas se ampliara. Para ello se realizaron pruebas en idiomas mayas. Cuando las tendencias en política pública buscaron establecer y sistematizar procesos de aseguramiento de la calidad, se elaboraron estándares y evaluaciones vinculadas a ellos. Para responder con agilidad y propiedad, las unidades de evaluación requieren personal capacitado, tiempo y apoyo financiero. La experiencia ha mostrado que si los encargados de la política educativa no están preparados para proveer los recursos necesarios, los proyectos de evaluación sufrirán.

Hubo una proliferación de sistemas de evaluación en América Latina en la década de 1990 (PREAL, 2001), siguiendo la tendencia de países industrializados. En estos últimos la fundación de los sistemas siguió una progresión natural, pero en varios países centroamericanos se establecieron como respuesta a demandas de una gestión gubernamental específica, de las condiciones establecidas por agencias de cooperación internacional o por condiciones de préstamos particulares. En Guatemala, por ejemplo, una de las razones por las que PRONERE se estableció fue responder a los requisitos de un préstamo recibido del Banco Mundial. Cuando el préstamo caducó, el sistema enfrentó retos financieros que se iniciaron en el año 2000 y no se solucionaron sino

[3] La división político-administrativa de Guatemala es el departamento. Cada departamento está compuesto de municipios.

hasta el año 2004. Entre los años 2004 y 2005 un nuevo sistema se estableció en respuesta a una iniciativa nacional. En esta ocasión el margen de tiempo provisto para la creación de la unidad de evaluación fue también de menos de un año. Plantear un sistema adecuado en un tiempo tan corto genera grandes presiones sobre las capacidades técnicas, organizacionales, financieras y operativas de todos los involucrados (Ferrer & Arregui, 2002).

Las políticas buscan resultados positivos, pero una hueste de factores adicionales influyen para que esto se concrete o no, en particular en sociedades donde hay grandes disparidades. La influencia de las políticas educativas en los resultados es compleja y también se encuentra constreñida por factores administrativos y presupuestarios. Un resultado óptimo requiere negociación y balance de las consideraciones técnicas (en este caso psicométricas), políticas y contextuales (Haddad & Demsky, 1995). Una comunicación asertiva puede prevenir que los encargados de la política pública tomen decisiones que afecten negativamente la calidad psicométrica de la evaluación.

Para los psicometristas, la evaluación es un medio de medir y monitorear los resultados educativos. Sin embargo, la evaluación también puede volverse parte de un proceso político por varias razones secundarias. Por ejemplo, grupos de presión pueden exigir cambios en la estructura de las evaluaciones sin consideración a los principios psicométricos. También son susceptibles a otras influencias, por ejemplo cuando los administradores operan cambios en los métodos de administración de las evaluaciones por razones puramente financieras. Estos son riesgos permanentes porque la evaluación tiene implicaciones presupuestarias, recibe atención de los medios de comunicación y pueden, incluso, convertirse en tema de debate nacional.

Validez, sesgo y equivalencia: las contribuciones de los expertos en el desarrollo de la evaluación educativa

La información de las evaluaciones educativas sólo es útil cuando refleja con exactitud lo que los estudiantes han aprendido. En otras palabras, es necesario que la evaluación sea válida. En contextos heterogéneos, los varios grupos que componen la sociedad con frecuencia tienen acceso diferenciado a los insumos educativos. Considérese la evaluación de la implementación de la EBI. El primer idioma de estos niños no es el español y se ubican principalmente en áreas rurales, donde hay un déficit de condiciones favorables. Desde una perspectiva se puede argumentar que las evaluaciones necesitan acomodarse a las diferencias poblacionales. Esto debido a que los estudiantes bilingües mayas de áreas rurales han sido sometidos a experiencias familiares y culturales diferentes, han tenido acceso a distintos recursos económicos y sus escuelas enfrentan retos particulares. Por otro lado, también puede argumentarse que las evaluaciones debieran

ser las mismas para todas las poblaciones. Esto debido a que la equidad exige que el sistema tome las medidas necesarias para asegurar que todos los estudiantes cuenten con las condiciones que les permitirán llenar los estándares, siendo la evaluación la única manera de detectar cuándo esto se ha logrando. He allí cómo la validez se ve afectada por las posibles interpretaciones que puedan darse a los resultados.

La definición tradicional de la validez de un instrumento es su capacidad para medir lo que pretende medir y se ha clasificado como predictiva (o referida al criterio), de contenido y de constructo (Crocker & Algina, 1986; Nunnally & Bernstein, 1995; Brualdi, 1999). El modelo predictivo se basa en la correlación entre los resultados de las evaluaciones y un criterio externo que es reflejo de la puntuación real del criterio. Los estudios de validez de contenido buscan establecer el vínculo entre los procedimientos utilizados para generar las puntuaciones de criterio y la interpretación propuesta, tomando como referencia una muestra de desempeños de alguna área de actividad, la cual es un estimador de la habilidad general. El modelo basado en el constructo ha evolucionado, pero actualmente alude a la acumulación de evidencia para elaborar una teoría que bosqueja la supuesta naturaleza del constructo o dominio. Esta forma de concebir la validez fue sometido a análisis crítico durante la década de 1980, cuando surgieron propuestas de un modelo unificado que tomara en cuenta las implicaciones de dar ciertos significados a las puntuaciones. Esta nueva perspectiva enfatiza la influencia que las interpretaciones tienen sobre las decisiones que se toman, las cuales conllevan consecuencias sociales (Messick, 1994; Brualdi, 1999; Kane, 2006).

Esta perspectiva, la cual tiene implicaciones sociales más amplias, ha ganado terreno, aun cuando todavía hay debates al respecto (e.g., ver Reckase, 1998 and Markus, 1998). Bajo esta concepción, la validez de constructo incluye la evidencia del constructo y reconoce el papel de los supuestos que subyacen a la interpretación de las puntuaciones (Kane, 2006). Estos supuestos deben ser evaluados con base a argumentos sólidos, dado que también pueden ser disputados. Se utilizan, entonces, seis criterios para corroborar la validez: (i) contenido, (ii) factores substantivos, (iii) estructura en relación con manifestaciones conductuales inherentes al constructo, (iv) capacidad de generalización, (v) sustentación externa de la consistencia entre la observación empírica y los significados asignados a las puntuaciones y (vi) consecuencias que resultan de la interpretación propuesta (Messick, 1994). Existen dos amenazas principales a la validez de una prueba (Brualdi, 1999). La primera es la subrepresentación del constructo, en donde la prueba no contiene una muestra representativa de las conductas que componen el constructo. La segunda es la varianza por factores que no son relevantes para el constructo. Esta se genera cuando los evaluados muestran variaciones en las puntuaciones que no pueden ser explicadas por sus diferencias en el constructo.

Esta concepción amplia de validez enfatiza la conexión que existe entre las políticas educativas y el desarrollo de las pruebas con las que se evalúa su impacto. Las

declaraciones de política educativa adhieren un valor a ciertas acciones y proponen una relación entre acciones y efectos conducentes a objetivos específicos. Por tanto, estas declaraciones debieran formar la base de las proposiciones y argumentos que subyacen a la construcción de las pruebas. El desarrollo de las pruebas se basa en el supuesto de que éstas proveerán información que es posible interpretar con sentido y que se vincularán a políticas específicas.

Las políticas que buscan favorecer la equidad requieren evaluación válida para poblaciones diferentes. Sin embargo, es muy probable que las diferencias que existen entre grupos se encuentren asociadas a variaciones que no tienen relación con el constructo. Por ejemplo, estudiantes con capacidades similares de áreas rurales y urbanas pueden obtener resultados diferentes en un ítem cuando el contexto del ítem se refiere a un mercado tradicional o a un centro comercial urbano. Para prevenir dichos efectos se pueden adaptar ítems para apegarse a las circunstancias locales. Es necesario, entonces, encontrar una forma de determinar si dicha adaptación fue exitosa.

La adaptación de evaluaciones (el desarrollo de evaluaciones para poblaciones heterogéneas) intenta que las pruebas sean válidas para todas las poblaciones efectuando análisis de comparación (o equivalencia) (Matthews-López, 2003). Para lidiar con la varianza no explicada por el constructo en medios heterogéneos (incluyendo heterogeneidad cultural) es necesario emplear dos conceptos esenciales: sesgo y equivalencia.

El sesgo es una condición psicométrica que ocurre cuando las diferencias en una medición no reflejan las diferencias del constructo entre examinados. Se han identificado tres tipos de sesgo (Van de Vijver & Poortinga, 2005; Van de Vijver & Tanzer, 2004). El sesgo de constructo ocurre cuando el constructo teórico es diferente, o no es equivalente, entre grupos. En ese caso no es posible obtener puntuaciones comparables o equivalentes. El sesgo por método involucra factores relacionados con la administración de las pruebas. El efecto de dichos factores usualmente puede ser detectado y eliminado o reducido, mejorando así la equivalencia de las puntuaciones para diferentes poblaciones (Van de Vijver & Leung, 1997). El sesgo de ítem o Funcionamiento Diferencial del Ítem (FDI) ocurre cuando hay anomalías en ítems específicos que causan diferencias en las puntuaciones de examinandos con capacidades similares. Por ejemplo, puede haber FDI cuando una traducción deficiente hace que la respuesta sea más obvia para estudiantes de ciertos grupos lingüísticos. El sesgo puede ser interno o externo (o predictivo) (Hong & Roznowski, 2001). El sesgo interno afecta la "justicia" de la prueba porque ocurre cuando dos estudiantes igualmente capaces tienen diferentes probabilidades de contestar correctamente. El sesgo externo o predictivo se da cuando la relación entre pruebas y el criterio externo difiere para cada grupo evaluado.

Un concepto relacionado al sesgo es el de la equivalencia. Se refiere al grado hasta el cual pueden hacerse comparaciones entre puntuaciones. El sesgo amenaza la

equivalencia al hacer difícil o imposible la comparación de resultados entre grupos. Para ser equivalentes, los instrumentos debieran medir el mismo constructo en cada grupo, exhibir la misma relación entre ítems (equivalencia estructural), compartir las mismas unidades de medición entre poblaciones (unidad de equivalencia métrica) y tener el mismo origen para todas las poblaciones (equivalencia completa o escalar) (Van de Vijver & Tanzer, 2004).

Es necesario que los resultados del análisis de sesgo y equivalencia de los instrumentos sean integrados al argumento interpretativo de la evaluación. Los sesgos asociados a la pertenencia a ciertas poblaciones tendrán un efecto en la distribución de las puntuaciones y, por tanto, debieran ser considerados en su interpretación. La interpretación se verá afectada por el tipo y extensión del sesgo identificado. En un extremo, cuando hay evidencia de sesgo en unos cuantos ítems, la solución puede ser eliminarlos. Pero en otro extremo, cuando no se logra demostrar la equivalencia de constructos entre pruebas, será necesario desarrollar instrumentos separados para cada grupo con el objeto de obtener una validez adecuada para cada uno. Es más frecuente encontrar sesgos moderados; la forma en la cual esto afectará la validez y las interpretaciones que de ellos se hagan depende de circunstancias cuyo análisis excede el alcance de este artículo. Hay que mencionar que mientras los conceptos de sesgo y equivalencia se utilizan con mayor frecuencia al discutir diferencias entre poblaciones culturales, el mismo tipo de análisis puede llevarse a cabo cuando existen otras características relevantes que dividen a la población en grupos (como, por ejemplo, el idioma hablado en casa o la localización urbano-rural de la comunidad).

En Guatemala la población maya sufrió por mucho tiempo la exclusión de varios servicios, tales como la educación. Actualmente, un alto porcentaje de la población asiste a escuelas rurales que no cuentan con insumos adecuados, como infraestructura y equipamiento (DP Tecnología, 2002; Esquivel Villegas, 2000). La interpretación de las puntuaciones de evaluaciones debe considerar el grado de influencia que dichas condiciones tienen sobre el desempeño. Otros análisis pueden explorar la manera en la cual la educación formal es aceptada por varios de estos grupos (e.g., ver Fuller & Clarke, 1994). La relevancia de estos factores en los resultados educativos debe ser evaluada para fortalecer la calidad educativa.

Diseminación de resultados y argumentos sobre validez

Una de las fases tardías en la evaluación educativa es la diseminación de información. En esta etapa la información es puesta a disposición de los actores relevantes, quienes la emplearán en formas diferentes. Tal como lo hizo el investigador, otros tomadores de decisiones, incluyendo a los encargados de las políticas, interpretarán los resultados

dentro de un marco de conocimientos, valores y creencias. Debido a ello es posible que se le den diferentes interpretaciones a un mismo resultado. Por ejemplo, si los estudiantes de una minoría obtienen una puntuación inferior al promedio nacional, algunos interpretarán los resultados como prueba de la exclusión a la que esa minoría está sometida. Otros podrían interpretar esos resultados como evidencia de que los estudiantes de ese grupo tienen una baja motivación o que no se esfuerzan lo suficiente. Por ello es importante que los hallazgos se encuentren acompañados de información clara sobre la validación de la evaluación, el riesgo del sesgo y las interpretaciones apropiadas a las que los datos pueden ser sometidos. La información del alcance y limitaciones del estudio debieran influir sobre las decisiones y recomendaciones que harán los encargados de las políticas educativas.

En una etapa previa se debe decidir cuáles serán las características de los estudiantes, docentes y escuelas que se analizarán. Después de todo, cada análisis requiere datos que deben ser colectados. El tipo de información que será recabada también requiere negociación con los encargados de políticas educativas. Así como es posible colectar información sobre factores de exclusión, también puede darse énfasis al estudio de las variables relacionadas con la efectividad escolar. La planificación de la evaluación escolar, así como la implementación de sugerencias basadas en sus hallazgos, requiere de negociación, primero entre encargados de políticas y expertos en evaluación, y más tarde con actores clave que tienen un rol en la recolección de datos o un interés particular en los resultados.

Los responsables de la evaluación deben comunicarle con claridad sus expectativas a las instituciones y personas que serán evaluadas, particularmente cuando éstas expectativas puedan tener efectos significativos sobre los procedimientos de administración de las evaluaciones. Lo mismo aplica para los trabajadores de campo que aplicarán las pruebas y los diseñadores de los instrumentos. Estas condiciones son bastante explícitas y cuando se les presta atención se logra prevenir en buena medida la aparición de algunas fuentes de sesgo.

Cuando la evaluación forma parte de un proceso general de aseguramiento de la calidad, la comunicación también puede afectar las decisiones de los encargados de las políticas educativas. Los sistemas de rendición de cuenta que funcionan de manera adecuada proveen información sobre los éxitos y limitaciones de las acciones, vinculan la evaluación con declaraciones oficiales de expectativas (estándares) utilizando métodos sistemáticos (mediciones de calidad, evaluación de desempeño) y crean condiciones para un adecuado seguimiento (Hanushek & Raymond, 2001). Los resultados de las evaluaciones se convierten en la medida del éxito o fracaso de la acción de política educativa. Comunicar resultados de manera inapropiada puede generar confusiones que afectarán negativamente las acciones. También puede dar lugar a la elaboración de

conclusiones que exceden los alcances de la evaluación, ya sea a favor o en oposición de las políticas bajo consideración.

Por ejemplo, si se distribuyen libros de texto que contienen un método novedoso para la enseñanza de la lectoescritura, es poco probable que se observen resultados dramáticos en el primer año. Además, los resultados aún sufrirán la influencia de otras variables, tales como el número de libros disponibles en el hogar. Los encargados de las políticas educativas, quienes usualmente son responsables por rendir cuentas, pueden no estar al tanto de las sutilezas técnicas de la evaluación, ocasionalmente dejando de tomar en cuenta estos efectos. Las diferencias de antecedentes académicos y prioridades profesionales entre expertos en evaluación y los responsables por las políticas educativas pueden crear fácilmente un abismo que les separará (Reimers & McGinn, 1997). Mientras que las preocupaciones de los psicometristas estarán orientadas por principios científicos, los encargados de las políticas educativas estarán interesados en factibilidad, recursos e impacto político (Reimers & McGinn, 1997). Por tanto, comunicar información de manera adecuada requiere reportar el alcance, limitaciones y posibles interpretaciones de los resultados. Esto no siempre ha recibido la atención adecuada. Un diagnóstico de las unidades de evaluación de la región latinoamericana encontró que en varios países el análisis y diseminación de los resultados no han recibido siempre la atención necesaria (Ferrer, 2006; Ferrer & Arregui, 2002; PREAL, 2001; Ravela, Wolfe, Valverde & Esquivel, 2006; Wolff, 2006).

Las políticas establecen metas y los medios para lograrlas. Tienen, además, referentes en un marco de valores. Por ejemplo, la promoción de la equidad presupone que ese estado es mejor que la segregación. Las políticas educativas podrían diseñarse para favorecer la segregación, aun cuando esto no sea considerado ético el día de hoy. Son los valores que hemos asumido los que conducen a aceptar la equidad y rechazar la segregación. Para que la investigación relacionada con políticas educativas sea relevante, tiene que ser consistente con las declaraciones de política. Si la política persigue la equidad, el tipo de investigación que hará sentido es la que investiga factores que favorecen la equidad y evitan o previenen las disparidades. Lo opuesto no es coherente. Es decir, restringir las políticas educativas a lo que los investigadores consideren importante estudiar no es adecuado, porque la investigación no provee un marco para valorizar y dar prioridad a una meta u objetivo sobre otro (Reimers & McGinn, 1997). La justificación que valida las investigaciones realizadas debiera ser consistente con el marco más amplio que razona la política educativa, así como debe ser consistente con el más específico que declaró las posibles interpretaciones de las puntuaciones de las pruebas. Si no se asegura dicha consistencia, será difícil que la evaluación pueda proveer realimentación adecuada al responsable de la política educativa.

En este sentido, los estudios de sesgo, equivalencia y validez en general son relevantes, pero no sólo como indicadores matemáticos de propiedades psicométricas.

Son relevantes especialmente como argumentos que clarifican la interpretación de los dominios estudiados. La evaluación es útil para las políticas cuando tiene propósitos claros, una filosofía orientada a la construcción de una visión de responsabilidad compartida, un diseño psicométrico de calidad, una orientación al apoyo de tareas docentes, respaldo de voluntad política para solucionar las deficiencias detectadas y capacidad para ubicar los indicadores numéricos en contexto (Ravela, Arregui, Valverde, Wolfe, Ferrer, Martínez Rizo, Aylwin & Wolff, 2008). La evaluación válida provee un marco de referencia útil para planificar soluciones para las deficiencias identificadas en el sistema educativo, cuando los resultados son comunicados efectivamente a aquellos que compartirán la responsabilidad. Por tanto, los argumentos que respaldan la validez de la evaluación deben ser accesibles a los actores relevantes locales (docentes, padres y madres y estudiantes). Con gran frecuencia éstos no tendrán conocimientos psicométricos. Aun así, son los ejecutores o detractores más directos de las políticas educativas. Si los argumentos no son claros y por tanto no les hacen sentido, aparecerán dificultades significativas cuando sea necesario traducir las políticas en actividades concretas.

Discusión y conclusiones

Recientemente se ha dado mucha atención y apoyo a la evaluación. Las demandas y expectativas de lo que se debiera lograr con ella probablemente aumentarán, potencialmente incrementando las tensiones entre los consumidores y productores de esa información. Para que la evaluación pueda contribuir a la calidad educativa, debe ser tanto factible como válida. Factibilidad implica aquí que la evaluación pueda llevarse a cabo logística y financieramente, al tiempo que conserva las características que la hacen relevante (e.g., especificaciones del muestreo, procedimientos de administración). La validez implica que la evaluación se basa en un argumento lógico consistente y que vincula las herramientas y procedimientos con los dominios bajo evaluación. No es sencillo lograr que las evaluaciones sean útiles para los encargados de las políticas educativas.

Las políticas influirán sobre la evaluación porque requerirán que los expertos desarrollen técnicas específicas para las necesidades de interés. Se presentaron varios ejemplos en la breve historia de la evaluación guatemalteca. Los estándares fueron creados para responder a un nuevo currículo en el cual las competencias se convirtieron en la expresión de las metas educativas. Fue necesario desarrollar instrumentos que tuvieran mayor representación del dominio, requiriendo múltiples formas y administración en distribución en espiral. Esto condujo a adoptar análisis basados en TRI. Lamentablemente, la influencia de los actores claves también puede ser menos constructiva. El caso de la resistencia ejercida por las escuelas privadas y su impacto en la muestra es un ejemplo.

La valorización e interpretación que los actores clave hacen de las puntuaciones también puede ser problemática. Los administradores de la política pública pueden tener prejuicios acerca de cómo funciona la evaluación y expectativas de lo que la misma podrá hacer. Esto no necesariamente corresponde a una práctica psicométrica adecuada. Sin embargo, son las políticas públicas las que aportan el marco de juicios y valores que dictan las acciones de un gobierno. Por tanto, los responsables de la evaluación deben distinguir entre los cambios en el diseño de la evaluación que afectarán la generalización (tal como el muestreo de los estratos) de aquellos que la invalidarán (como administrar herramientas que no han sido piloteadas de forma adecuada). Al tener claridad en esto, cuando el evaluador deba discutir recursos, márgenes de tiempo, condiciones de administración y otras, sabrá qué aspectos son negociables cuando se le solicite un cambio en la evaluación y cuáles no lo son dado que invalidarían el estudio. El resultado de estas negociaciones podría alterar el alcance o envergadura de la evaluación, pero no debiera alterar su perspectiva científica. Esto también tiene implicaciones sobre el uso de la información durante la etapa de seguimiento. Los encargados de las políticas no siempre comprenderán que es necesario cambiar los usos que se dará a los resultados como consecuencia de los cambios hechos para asegurar la factibilidad. Por ejemplo, debido a las restricciones financieras puede haber sido necesario reducir el número de formas de las pruebas que se utilizaron. Esto reduce el número de ítems administrados, por lo que no podrán darse resultados por subárea. En este caso se entregaría una puntuación total para matemática, pero no una para sumas, restas, geometría, etc. Sin embargo, el encargado de la política pública puede encontrar atractivo proveer resultados por subárea, aun cuando se han aplicado sólo uno o dos ítems para cada una de ellas. Es el reto de los evaluadores prevenir esto, comunicando de manera apropiada al encargado de la política las limitaciones y consecuencias de dicha acción. De otra manera, se planificarán intervenciones educativas con base a resultados que no son válidos.

La evaluación también es vulnerable a las influencias de los grupos de presión y a los intereses de cada gestión gubernamental sobre la continuidad de procesos. Esto sucede cuando se intenta inapropiadamente utilizar la evaluación como herramienta de intervención, en lugar de utilizarla para el monitoreo; o cuando la evaluación se convierte en parte de un modelo de rendición de cuentas que es impulsado por la política pública. Nuevamente, el encargado de políticas públicas que tiene insuficiente información psicométrica puede tomar decisiones contraproducentes. Por otra parte, una visión de los temas técnicos que no considera los antecedentes políticos de las decisiones puede afectar negativamente la factibilidad y continuidad del proceso de evaluación. En particular, ignorar a los actores clave locales en el esfuerzo debilitará el impacto de la evaluación al nivel local.

Las perspectivas actuales consideran que la validación puede asumirse como la verificación de una hipótesis, tomando también en cuenta los marcos de referencia de

valores. Estas posiciones sugieren que la validación debe verse como un argumento que consolida evidencia e intenta contrarrestar las fuentes de invalidez (ver Kane, 2006; Messick, 1994, 1998). El investigador debiera establecer los modos y canales para comunicar de manera clara el razonamiento que da cuerpo a la evaluación y a su generalización. Esto sólo puede hacerse por medio de una perspectiva común que vincula la explicación de los efectos de cierta política sobre objetivos concretos, con el razonamiento que explica cómo una evaluación específica refleja los dominios evaluados, con la justificación que torna las puntuaciones de esa evaluación en indicadores, con las posibles interpretaciones que se pueden hacer de los datos. Si no se sigue este camino, las medidas de política estarán desarticuladas de sus indicadores de éxito o fracaso, los cuales provienen de la evaluación. Bajo dicho escenario, los actores clave locales también tendrán dificultades para vincularse a la política nacional. Es necesario asegurar la consistencia de perspectivas en la política, evaluación y acciones locales.

Estas ideas y temas no son nuevos y pueden parecer autoevidentes. Sin embargo, recalcarlos está en orden. No es inusual en muchos países latinoamericanos encontrar una falta de continuidad y consistencia entre los esfuerzos de política y los técnicos. La continuidad no puede definirse sólo como la adopción de declaraciones de políticas que sintácticamente son similares y trascienden una gestión en particular. La continuidad está dada por la conservación consistente del razonamiento y argumentación que sustentan a las políticas educativas, las estrategias de evaluación de su logro y la implementación local. En ningún otro lugar es esto más relevante que en el sector educativo, en donde el éxito se basa en procesos que consumen tiempo y demandan esfuerzos consistentes y sistemáticos.

Bibliografía

Álvarez, H. & Schiefelbein, E. (2007, diciembre). *Informe integrado del sector educación: Informe final.* (Reporte financiado por el Banco Interamericano de Desarrollo y la Agencia Sueca para el Desarrollo Internacional como insumos para la estrategia en Armonización y Alineación de Políticas). Guatemala: Ministerio de Educación (BID / MINEDUC / ASDI).

Antillón Milla, J. (1997). La educación. En Asociación Amigos del País (Ed.) *Historia General de Guatemala: Vol. VI. Época contemporánea, de 1945 a la actualidad* (pp. 591-612). Guatemala: Asociación de Amigos del País, Fundación para la Cultura y el Desarrollo.

Beckett, M. & Pebley, A. R. (2003). Ethnicity, language, and economic well-being in rural Guatemala. *Rural sociology,* 68, 3, 434-458.

Brualdi, A. (1999). *Traditional and modern concepts of validity* (Research Report No. RR-94-95). Washington, DC: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service Nº ED435714).

Buckendahl, Ch. W.; Smith, R. W.; Impara, J. C. & Plake, B. S. (2000, October). *A comparison of the Angoff and Bookmark standard setting methods.* Paper presented at the annual meeting of the Mid-Western Educational Research Association in Chicago, IL.

CCRE (Comisión Consultiva para la Reforma Educativa. (1996). *Diseño de reforma educativa.* Guatemala: Ministerio de Educación.

Chesterfield, R.; Rubio, F. E.; Vásquez, R. (2003, April). *Study of bilingual education graduates in Guatemala.* Guatemala: Juárez & Associates, Inc., MEDIR Project. (Prepared under task order to the Improving of Educational Quality II, Project # 536-5858, USAID/G-CAP).

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory.* Orlando, FL: Harcourt Jovanovich College Publishers.

de Baessa, Y. (1997). Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: 1997. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 1997 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (1998). Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar de 1998. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 1998 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (1999a). Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: 1999. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 1999 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (1999b). Informe sobre los Resultados de la Aplicación de Pruebas en Idiomas Mayas en Tercer Grado Primaria: 1999. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 1999 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (2000a). Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: Año 2000. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 2000 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (2000b). Informe sobre los Resultados de la Aplicación de Pruebas en Idiomas Mayas en Tercer Grado Primaria: Diciembre 2000. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 2000 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (2001a). *Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: Año 2001* (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 2001 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

de Baessa, Y. (2001b). Informe sobre los Resultados de la Aplicación de Pruebas en Idiomas Mayas en Tercer Grado Primaria: Diciembre 2001. (Reporte de la Universidad del Valle de Guatemala de la evaluación del año 2001 bajo contrato con el Ministerio de Educación de Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministerio de Educación.

Dings, J.; Childs, R. & Kingston, N. (2002, January). *The effects of matrix sampling on student score comparability in constructed-response and multiple-choice assessments.* Washington, D.C.: State Collaborative on Assessment and Student Standards (SCASS).

DP Tecnología, S.A. (2002, February). Estudio Cuasi-Experimental de Resultados de PRONADE Año 2001. (Reporte de DP Tecnología bajo contrato con el Ministerio de Educación). Guatemala: DP Tecnología / Ministerio de Educación.

Esquivel Villegas, F. (2006, February). *Situación del sistema educativo guatemalteco.* (Reporte de consultoría desarrollado para el Ministerio de Educación y el Banco Mundial en preparación para la negociación de un préstamo al gobierno de Guatemala). Guatemala: Ministerio de Educación de Guatemala y Oficina Nacional para el Banco Mundial.

Ethnologue, languages of the world: Languages of Guatemala (2005, April). Retrieved March 10, 2008, from http://www.ethnologue.com

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges.* USA: Partnership for Educational Revitalization in the Americas (PREAL).

Ferrer, G. & Arregui, P. (2002, October). La experiencia latinoamericana con pruebas internacionales de aprendizaje: Impacto sobre los procesos de mejoramiento de la calidad de la educación y criterios para guiar las decisiones sobre nuevas aplicaciones. Lima: GRADE (Grupo de Análisis para el Desarrollo).

Fuller, B. & Clarke, P. (1994). Raising school effect while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research,* 64 (1), 119-157.

Gobierno de Guatemala, Unidad Revolucionaria Nacional Guatemalteca & Naciones Unidas (1996). *Acuerdo de paz firme y duradera*. Extraído en abril 1 del 2008 de www.congreso.gob.gt/Docs/PAZ/Acuerdo%20de%20paz%20firme%20y%20duradera.pdf

Haddad, W. D. & Demsky, T. (1995). *Educational policy-planning process: An applied framework.* Paris: UNESCO, International Institute for Educational Planning.

Hanushek, E. A. & Raymond, M. E. (2001). The confusing world of educational accountability. *National Tax Journal,* 54 (2), 365-384.

Hong, S. & Roznowski, M. (2001). An investigation of the influence of internal test bias on regression slope. *Applied Measurement in Education,* 14 (4), 351-368.

Jiménez Sánchez, A. O. (1998, September). *Mayan languages and the Mayan movement in Guatemala.* Chicago, Illinois: Latin American Studies Association. Retrieved April 9, 2008: http://www.tamilnation.org/selfdetermination/countrystudies/mayan.pdf

Kane, M. T. (2006). Validity. In R. L. Brennan (Eds.), *Educational measurement, ongoing themes in psychology and culture* (pp. 207-222). Westport, CT: Praeger Publishers.

Kiplinger, V. L. (1997, October). Performance levels on a standards-based assessment. Colorado: Colorado Department of Education. Retrieved January 9, 2006: http://www.cde.state.co.us/cdeassess/csap/asperf.htm

MacCann, R. G. & Gordon, S. (2004, January). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical assessment, research & evaluation.* Chicago, Illinois: Latin American Studies Association. Retrieved January 9 of 2006 from: http://PAREonline.net/getvn.asp?v=9&n=5

Markus, K. A. (1998). Validity, facts, and values sans closure: Reply to Messick, Reckase, Moss, and Zimmerman. *Social Indicators Research,* 45, 73-82.

Matthews-López, J. L. (2003). *Best practices and technical issues in cross-lingual, cross-cultural assessments: An evaluation of a test adaptation.* Unpublished doctoral dissertation, College of Education of Ohio University - Ohio, USA.

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (Research Report No. RR-94-95). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service Nº ED380496).

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research,* 45, 35-44.

Ministerio de Educación de Guatemala (2005, noviembre). *El nuevo currículo, su orientación y aplicación.* Guatemala: Ministerio de Educación de Guatemala.

Ministerio de Educación de Guatemala & Programa Estándares e Investigación Educativa-USAID (2006, December). *Estándares Educativos de Guatemala.* Guatemala: Ministerio de Educación - USAID.

Moreno Grajeda, M.R., Gálvez-Sobral, J.A., Bedregal, S. & Roldán, K. (2008, April). Informe de Resultados, Evaluación de Primaria 2006, Tercer Grado. (Reporte de la Unidad para Análisis Estadístico de DIGEDUCA del Ministerio de Educación / Universidad del Valle de Guatemala / Ministerio de Educación.

Nunnally, J. C. & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). USA: McGraw-Hill Series in Psychology.

Porta, E. & Laguna, J. R. (2007). *Guatemala, country case study* (Paper commissioned for the EFA Global Monitoring Report 2008, Education for All by 2015: will we make it?). Retrieved April 9, 2008, from http://unesdoc.unesco.org/images/0015/001555/155575e.pdf

Porta, E. & Somerville, S. (2006). *Sistema nacional de indicadores educativos, Guatemala.* Guatemala: USAID. Extraído en abril 9 del 2008, de http://www.mec.es/educa/rieja/files/guatemala-sistema-nacional-indicadores-educativos.pdf

PREAL (2001). ¿Cómo avanzar en la evaluación de aprendizajes en América Latina? *Formas & Reformas de la Educación, Serie Políticas,* 2 (8).

Ravela, P.; Wolfe, R.; Valverde, G. & Esquivel, J. M. (2006). Los próximos pasos: ¿Cómo avanzar en la evaluación de aprendizajes en América Latina? In P. Arregui (Ed.), *Sobre estándares y evaluaciones en América Latina* (pp. 52-121).

Reckase, M. D. (1998). The interaction of values and validity assessment: Does a test's level of validity depend on a researcher's values? *Social Indicators Research,* 45, 45-54.

Reimers, F. & McGinn, N. (1997). *Informed dialogue: Using research to shape education policy around the world.* CT: Praeger.

Richards, M. (2003). *Atlas lingüístico de Guatemala.* Guatemala: SEPAZ, UVG, URL, USAID.

United Nations Development Programme (2006). *Human development report 2006; Beyond scarcity: Power, poverty and the global water crisis.* New York: United Nations Development Programme.

Van de Vijver, F. and Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry, Y.H. Poortinga, and J. Pandey (Eds.). *Handbook of cross-cultural psychology, volume 1: Theory and method* (2nd ed.). Boston: Allyn and Bacon.

Van de Vijver, F. J. R. & Poortinga, I. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda & Ch. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Van de Vijver, F. & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée,* 54, 119-135.

Wolff, L. (2006). Las evaluaciones educacionales en América Latina: Avance actual y futuros desafíos. En P. Arregui (Ed.), *Sobre estándares y evaluaciones en América Latina* (pp. 13-52).

World factbook. (2007). USA: CIA. Retrieved May 27, 2008, from https://www.cia.gov/library/publications/the-world-factbook

FECHA DE RECEPCIÓN: 5 de octubre de 2008 FECHA DE ACEPTACIÓN: 15 de octubre de 2008