Debates en Evaluación y Currículum
Posgrado Educación UATx
Congreso Internacional de Educación
26, 27 y 28 de septiembre

ISSN: 2448-6574

Experiencias en la conformación de pruebas verticales para la evaluación de aprendizajes de estudiantes del ciclo básico de secundaria en Guatemala[1]

Francisco José Ureta Morales
fureta@mineduc.gob.gt

PRONACOM para la DIGEDUCA

Escuela de Ciencias Psicológicas y Facultad de Humanidades
Universidad de San Carlos de Guatemala, Guatemala

Evaluación curricular, acreditación de programas e impacto de las acciones de evaluación en el currículo.

Resumen.

La Dirección General de Evaluación e Investigación Educativa (DIGEDUCA) del Ministerio de Educación de Guatemala (MINEDUC), es la encargada de evaluar el aprendizaje de los estudiantes en todos los niveles educativos atendidos. Debido a deficiencias en el nivel básico de la educación básica secundaria, se implementa un programa de mejoramiento y formación de docentes. Se desarrollan pruebas verticales y matriciales en las áreas de comunicación y lenguaje, matemática y ciencias naturales para primero, segundo y tercero básico, que evaluarán el impacto de dicha intervención, denominadas AVANZO. Dicho proceso incluyó tres pilotajes de las pruebas y su aplicación en la línea de base de 2018, se realizaron los análisis de ítems y se conformaron las versiones finales de dichos instrumentos. En general tienen buenos índices Alfa de Cronbach de confiabilidad y las formas presentan dificultades equilibradas a los estudiantes. Con estos resultados se construyeron las formas finales de las escalas verticales y matriciales, para aplicar en las líneas intermedia y final del programa.

Palabras clave: evaluación del aprendizaje, pruebas verticales.

[1] Correspondencia dirigirla a Francisco José Ureta, Departamento de Análisis, Avenida Reforma 8-60 zona 9, Edificio Galerías Reforma, Torre II, 8º nivel, Ciudad de Guatemala, Guatemala. E-mail: fureta@mineduc.gob.gt

ISSN: 2448-6574

Planteamiento del problema.

La educación secundaria en Guatemala tiene algunas deficiencias en sus indicadores de contexto, insumos, proceso y productos, tal como puede observarse en el cuadro 1, donde aparecen los indicadores de los departamentos seleccionados para la intervención del Programa Umbral. Como una posibilidad de solución de esta problemática surge la iniciativa de la intervención Éxito Educativo para la educación secundaria, se concretó en un convenio que da origen al Programa Umbral. El cual fue ubicado en el Programa Nacional de Competitividad (PRONACOM) y a partir de esto se inició la ejecución del plan operativo y financiero del Programa en octubre de 2016. El ocho de abril de 2015 la República de Guatemala, a través del Ministerio de Economía, (MINECO), firmó el Convenio de Donación del Programa Umbral, con el Gobierno de los Estados Unidos de América, actuando a través de la Millennium Challenge Corporation (Corporación Desafío del Milenio, «MCC», por sus siglas en inglés). El convenio de donación del Programa Umbral en el área de educación, tiene como objetivo apoyar las reformas impulsadas por el Gobierno para mejorar la calidad y relevancia del nivel de educación media en Guatemala.

Cuadro 1

Indicadores de los 5 departamentos escogidos para el Programa Umbral

Indicadores educativos del ciclo básico	Departamentos a intervenir				
	Alta Verapaz	Sololá	Sacatepéquez	Jalapa	Chiquimula
Tasa de pobreza rural	89.6%	84.5%	80.5%	77.3%	79%
Nivel de logro pruebas de tercero básico, lectura	7.88%	5.65%	22.4%	6.7%	11.9%
Nivel de logro pruebas de tercero básico, matemática	12.2%	9.91%	25%	10.5%	15.1%
Tasas neta de escolaridad	26.3%	41%	60.9%	35.1%	34.2%
Tasa de retención	95.8%	94.6%	94.2%	92.7%	90.4%
Tasa de deserción	4.15%	5.33%	5.74%	7.24%	9.53%
Tasa de aprobación	72.7%	69.5%	60.7%	76.7%	78.4%
Tasa de reprobación	27.2%	30.4%	39.3%	23.2%	21.5%

Fuentes: (Ministerio de Educación, 2016), (Instituto Nacional de Estadística, 2011) y (Dirección General de Investigación y Evaluación Educativa, 2014).

ISSN: 2448-6574

Justificación.

La participación de Dirección General de Evaluación e Investigación Educativa (Digeduca) del MINEDUC es la responsable de desarrollar evaluaciones para conocer el nivel de la calidad de la educación de manera objetiva, confiable y válida, por medio de la aplicación de evaluaciones estandarizadas que faciliten la identificación de los aspectos que elevan o no la calidad en determinados momentos del proceso educativo. La evaluación del Nivel de Educación Media implementado tiene como objetivo evaluar los logros alcanzados por los estudiantes en las competencias establecidas para comunicación y lenguaje, matemática y ciencias naturales de primero, segundo y tercero básico, mediante evaluaciones estandarizadas. Este proceso evaluativo optó por construir una escala vertical con pruebas matriciales, la cual fue piloteada en 3 operativos con institutos y escuelas públicos y, colegios privados. Los cuales incluyeron un total de 207 establecimientos educativos y un total de 22384 estudiantes, situación que permitió preparar las pruebas para la aplicación de la línea base del proyecto en los meses de mayo y junio del 2018. En la línea de base se evaluó un total de 11,748 estudiantes de primero básico, en 331 institutos que incluyeron 158 de intervención y 173 de control. Se evaluaron las áreas de comunicación y lenguaje, matemática y ciencias naturales, así como la aplicación de cuestionarios de factores asociados para estudiantes y docentes que facilitan las áreas evaluadas en los estudiantes.

Fundamentación teórica.

Una parte integral de la construcción de los instrumentos fueron los análisis de ítems, que se realizaron con teoría clásica y teoría de respuesta al ítem, para la consistencia interna total se utilizó el Alfa de Cronbach, mientras más homogéneos sean los ítems, mayor será el valor de la consistencia interna para un número dado de ítems, referido por Magnusson en 1978, citado por (Quero Virla, 2010). Para evaluar la confiabilidad o la homogeneidad de las preguntas el coeficiente alfa de Cronbach analiza alternativas de respuestas politómicas, cuyos resultados toman valores entre 0 y 1, los resultados se entienden así: 0 significa confiabilidad nula y 1 representa confiabilidad total, (Corral, 2009). Para cada ítem, el Alfa de Cronbach por ítem se considera como criterio básico un resultado apropiado entre 0.8 a 1. (George & Mallery, 2003).

Para el análisis de la teoría de respuesta al ítem se aplicó el modelo de medida que en 1960 el matemático danés Georg Rasch propuso, el cual permite solventar muchas de las deficiencias

ISSN: 2448-6574

de la teoría clásica y construir pruebas más adecuadas y eficientes. El modelo propuesto por Rasch se fundamenta en los siguientes supuestos:

- El atributo que se desea medir puede representarse en una única dimensión en la que se situarían conjuntamente las personas y los ítems.
- El nivel de habilidad de la persona en el atributo y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta. Si el control de la situación es adecuado, esta expectativa es razonable y así debe representarla el modelo matemático elegido. (Prieto & Delgado, 2003).

Se trabajó con el modelo dicotómico de Rasch, correcto-incorrecto, el primer paso para la estimación de los valores de habilidad y dificultad del ítem es el cálculo del porcentaje de respuestas correctas para cada estudiante (número de respuestas correctas dividido por el total de respuestas) y para cada ítem (número de personas que acertaron el ítem dividido por el total de personas). Esto proporciona los porcentajes brutos de la prueba, los cuales son suficientes para calcular los parámetros de habilidad y dificultad. (Morales, Moreno, & Santos, 2015).

Para realizar el análisis de Rasch se utilizó el programa Winsteps, el cual posee la función de convertir las medidas de habilidad y dificultad del ítem en lógitos (expresan en este programa las estimaciones de habilidad y dificultad de los ítems) a otras medidas significativas. Para este proceso se utilizó el estadístico measure que refiere la habilidad del respondiente, el cual se normalizó a un promedio de 500 y una desviación estándar de 50, como una medida alternativa al porcentaje de respuestas correctas. Este programa comienza haciendo un estimado central para cada calibración o estimación de los parámetros de habilidad de los individuos y dificultad de cada ítem. Este estimado inicial da una aproximación al patrón de datos que se está observando. En un siguiente momento Winsteps aplica el procedimiento UCON (Estimación de Máxima Verosimilitud) para obtener mayor exactitud en las estimaciones de habilidad y dificultad del ítem. La estimación de ajuste entre datos y modelos se obtiene calculando los residuales entre cada individuo al responder a cada ítem. Este cálculo da un estimado de cuánto se apartan los patrones de respuesta de las expectativas del modelo. Los resultados de este cálculo se reportan en lógitos y estadígrafos de ajuste. En el programa Winsteps, estos estadígrafos de ajuste se llaman valores infit y outfit, tanto para los ítems como para los individuos según González en el 2008 (Morales, Moreno, & Santos, 2015).

ISSN: 2448-6574

El infit es un estadígrafo que captura comportamientos de respuestas no esperadas o anomalías a ítems calibrados cerca del nivel de habilidad del individuo. Un valor infit de 1.00, indica un ajuste perfecto entre los datos y el modelo. Valores superiores al 1.5 indican una falta de ajuste, así como una alta variabilidad aleatoria o ruido en los datos. Los valores menores a 1.00 también indican una falta de ajuste. El outfit es el estadígrafo sensible a los valores extremos y a comportamientos no esperados que afectan las respuestas a ítems que se encuentran lejos del nivel de habilidad del individuo. Un valor outfit de 1.00 indica un ajuste perfecto. Un valor superior a 1.5 indica una falta de ajuste ya que muestra presencia de valores extremos. Al igual que con el estadístico de Infit, los valores menores a 1.00 también indican una falta de ajuste, ya que estos no logran ajustarse al modelo. Son aceptables los ítems con valores superiores a 0.25, (Pardo, Rocha, Avendaño, & Barrera, 2005).

Tanto en Infit como en Outfit aparecen valores estandarizados llamados ZSTD. Cuando este estadígrafo toma valores de -2 a +2, entonces los valores están en el intervalo de lógitos aceptables para determinar el ajuste, tanto en individuos como en ítems. Los valores superiores a +2 e inferiores a -2 indican una falta de ajuste entre los datos y el modelo, indicado por González en el 2008 (Morales, Moreno, & Santos, 2015), así como por (Pardo, Rocha, Avendaño, & Barrera, 2005). La estimación de la dificultad de ítem, la cual tiene los valores que oscilan entre -3 ítem muy fácil hasta 3, ítem muy difíciles, 0 es la media, (Pardo, Rocha, Avendaño, & Barrera, 2005). Para estimar el punto biserial de cada ítem, se utilizó el criterio de valores por encima de 0.25 para incluirlos en el análisis y reporte de resultados finales, (Pardo, Rocha, Avendaño, & Barrera, 2005).

Objetivo.

Evaluar las áreas de comunicación y lenguaje, matemática y ciencias naturales con pruebas verticales y matriciales y, cuestionarios de contexto; a estudiantes de primero y tercero básico de una muestra de institutos por intervenir y control de los 5 departamentos de influencia del Programa Umbral.

ISSN: 2448-6574

Metodología.

Proceso para la definición y construcción de las Pruebas AVANZO

En el proceso de construcción de las pruebas verticales (las pruebas son diseñadas para evaluar un mismo constructo o concepto con diferentes niveles de dificultad, su propósito es la equiparación de pruebas que miden la misma habilidad pero en un amplio rango de dificultad o tiempo, en este caso varios grados, de cuarto primaria a tercero básico, (Pacheco–Villamil, 2007) y matriciales (su desarrollo involucra la construcción de un gran número de reactivos para cubrir la mayor cantidad de contenidos curriculares, por lo que un estudiante determinado responde solamente a un subconjunto de ellos, (Instituto Nacional para la Evaluación de la Educación, 2006) se siguió la siguiente secuencia de actividades:

- El referente de la prueba desarrollada fue el Currículo Nacional Base (CNB) desde tercero primaria a tercero básico vigentes, las competencias, indicadores de logro y contenidos.
- Realización de un mapeo curricular se concretó en una tabla de alcances y secuencias, la cual sirvió para la construcción de ítems que conformaron las pruebas, evidenciaron el progreso en el alcance de una competencia en un área de estudio.
- El análisis curricular realizado se centró en dos componentes curriculares, las competencias de las áreas y los indicadores de cada área evaluada, de cuarto primaria a tercero básico. Se indicó que la competencia o el indicador estaban cubiertos cuando al menos, un ítem cubría dichos elementos curriculares. En algunos casos fueron solamente un ítem, pero en otros, la mayoría, hubo más de un ítem, por lo que hay que indicar que tanto las competencias como los indicadores están cubiertos de manera parcial. Los ítems no evalúan toda la complejidad de las competencias, así como tampoco todos los indicadores en su justa dimensión. El formato de ítem de opción múltiple tiene esa limitante, además hay que recordar que se incluyeron aquellos indicadores y contenidos que tenían una progresión de grado a grado para el aprendizaje de los estudiantes.
- Considerando estos aspectos, se puede indicar que se cubren el 72.02% de las competencias y el 46.13% de los indicadores de logro, ofrecen una evaluación de casi el 50% del CNB de primaria y de los básicos en su revisión del 2017-2018. El área que

ISSN: 2448-6574

menos abarca es comunicación y lenguaje, ya que solamente se evalúa la comprensión lectora dejando de lado escuchar, hablar y escribir como áreas curriculares. Las áreas de matemática y ciencias naturales tienen mejores porcentajes de cobertura, en algunos casos se cubre parcialmente el 100% de competencias y hasta el 81.25% de los indicadores. Por lo tanto, se puede afirmar que los ítems que reflejaron consistentes características psicométricas para permanecer en la versión final de las escalas verticales, evalúan de manera apropiada cerca de la mitad de indicadores de logro, dado el formato de ítems de selección múltiple, que tienen sus bondades y limitaciones.

- Diseño de las tablas de especificaciones de las áreas a evaluar.
- Elaboración de los ítems por parte de consultores especializados en cada área, 900 por cada una de las áreas a evaluar y análisis de ítems de los tres pilotajes.
- Construcción de las pruebas finales para la línea de base.
- Aplicación de la línea de base con estudiantes de primero y tercero básico, en las áreas de comunicación y lenguaje, matemática y ciencias naturales.
- Análisis de ítems y cálculo del porcentaje de respuestas correctas.

Resultados análisis de ítems

Al concluir la aplicación y posterior digitación y lectura de las hojas de respuesta, se realizó durante agosto, septiembre y octubre del 2018 el análisis de los ítems de las pruebas de primero y tercero básico, el cual consideró los siguientes pasos: Calificación de las preguntas de cada forma de las áreas evaluadas, proceso realizado en el programa SPSS. Se calificó cada forma y área por separado y se calculó el índice Alfa de confiabilidad. El siguiente paso fue el análisis con teoría de respuesta al ítem con el programa Winsteps. Luego se analizó con el programa JMetrik. Esto permitió que al pedir los análisis se obtuvieran los resultados de los distractores para cada ítem, elemento importante para la revisión de los ítems realizado. El criterio es que cada distractor debe tener como mínimo un 5% de respuestas, para considerarlo un buen distractor, al no tenerlo se sugiere su remoción o cambio. Luego de ponerse de acuerdo los técnicos con sus análisis, se procedió a recalcular de nuevo el proceso sin los ítems eliminados. Procesos realizados por separado por dos analistas, para finalmente contrastar resultados y comprobar que los datos coinciden. Uno de los aspectos más importantes de los análisis de ítems fue el cálculo del coeficiente Alfa de Cronbach para la confiabilidad de las

ISSN: 2448-6574

pruebas, en general las formas de comunicación y lenguaje y ciencias naturales obtuvieron una confiabilidad apropiada. Las pruebas de matemática obtuvieron índices más bajos, por lo que se redujo la confiabilidad de las mismas, aun así, los instrumentos cuentan con buenas confiabilidades. Los siguientes cuadros así lo evidencian con los resultados eliminando los ítems con bajas propiedades psicométricas.

Cuadro 2

Índices alfa de Cronbach de la confiabilidad de las pruebas AVANZO, línea base primero básico

Área evaluada	Forma de la prueba	Índice de confiabilidad	Número de ítems	Estudiantes evaluados
Comunicación y lenguaje	2	0.9022	49	1,921
	3	0.9024	48	1,993
	4	0.9036	49	1,914
	5	0.9035	48	1,995
Matemática	1	0.736	41	2,020
	2	0.722	40	1,922
	4	0.714	40	1,913
	6	0.752	41	1,905
Ciencias naturales	1	0.852	45	2,020
	3	0.833	45	1,993
	5	0.854	44	1,995
	6	0.846	45	1,905

Fuente: procesos internos de Digeduca / MCC, 2017-2018.

Cuadro 3

Índices alfa de Cronbach de la confiabilidad de las pruebas AVANZO, línea base tercero básico

Área evaluada	Forma de la prueba	Índice de confiabilidad	Número de ítems	Estudiantes evaluados
Comunicación y lenguaje	8	0.880	48	1,006
	9	0.887	48	1,059
	10	0.878	49	1,018
	11	0.879	48	1,088
Matemática	7	0.725	45	1,073
	8	0.812	44	1,006

ISSN: 2448-6574

	10	0.751	44	1,018
	12	0.807	45	1,019
Ciencias naturales	7	0.866	46	1,073
	9	0.800	48	1,059
	11	0.854	48	1,088
	12	0.793	45	1,019

Fuente: procesos internos de Digeduca / MCC, 2017-2018.

A manera de conclusiones.

Para verificar que las formas ofrecieron las mismas dificultades a todos los estudiantes, se hizo la comparación entre formas para cada área evaluada, se calculó una anova para la comparación de los promedios de la dificultad de cada ítem en las 4 formas. Se tomó como dificultad la variable measure que calcula el programa Winsteps al realizar al análisis TRI con Rasch para cada ítem, la base se conformó con la dificultad de cada uno. En el caso de comunicación y lenguaje se obtuvo una F= 0.034, 3 grados de libertad inter grupos, p de 0.991 > 0.05, lo cual indica que los promedios son iguales entre las dificultades de las 4 formas con un 95% de probabilidad. Se calculó un estadístico más robusto con la prueba post hoc de Bonferroni, p de 1.0 > 0.05, este resultado indicó que todos los promedios son iguales entre las dificultades de las cuatro formas, con un 95% de probabilidad. En los casos de ciencias naturales y matemática los resultados fueron también de igualdad entre las 4 formas de cada área. En matemática se obtuvo una F= 0.037, 3 grados de libertad inter grupos, p de 0.991 > 0.05 y en la prueba post hoc de Bonferroni, p de 1.0 > 0.05. Para ciencias naturales se obtuvo una F= 0.002, 3 grados de libertad inter grupos, p de 1.000 > 0.05, la prueba post hoc de Bonferroni, p de 1.0 > 0.05. Resultados que confirman que las pruebas matriciales ofrecieron la misma dificultad en ítems a todos los estudiantes, con lo que se afirma que los resultados que se presentaron no fueron influidos por la forma que tomó cada estudiante evaluado de primero básico, (Dirección General de Evaluación e Investigación Educativa, 2019).

En el caso de tercero básico, comunicación y lenguaje se obtuvo una F= 1.02, 3 grados de libertad inter grupos, p de 0.384 > 0.05, lo cual indica que los promedios son iguales entre las dificultades de las 4 formas con un 95% de probabilidad. También se calculó el estadístico más robusto con la prueba post hoc de Bonferroni, p de 1.0 > 0.05, este resultado indicó que todos

ISSN: 2448-6574

los promedios son iguales entre las dificultades de las cuatro formas, con un 95% de probabilidad. En los casos de ciencias naturales y matemática los resultados fueron también de igualdad entre las 4 formas de cada área. En matemática se obtuvo una F= 0.000, 3 grados de libertad inter grupos, p de 1.0 > 0.05 y en la prueba post hoc de Bonferroni, p de 1.0 > 0.05. Para ciencias naturales se obtuvo una F= 0.035, 3 grados de libertad inter grupos, p de 1.000 > 0.05, la prueba post hoc de Bonferroni, p de 1.0 > 0.05. Resultados que confirman que las pruebas matriciales ofrecieron la misma dificultad en ítems a todos los estudiantes, con lo que se afirma que los resultados que obtenidos no fueron influidos por la forma que tomó cada estudiante evaluado en tercero básico.

Finalmente, los pilotajes y análisis permitieron identificar los ítems más robustos por sus características psicométricas, para las pruebas de primero básico, en comunicación y lenguaje tienen dificultades 3 ítems de 77 aplicados, el 3.89%. En ciencias naturales tienen dificultades 6 ítems de 66 aplicados, el 9.09%. En matemática tienen dificultades 12 ítems de 70 aplicados, el 17.14. Para las pruebas de tercero básico, en comunicación y lenguaje tienen dificultades 8 ítems de 76 aplicados, el 10.52%. En ciencias naturales tienen dificultades 12 ítems de 65 aplicados, el 18.46%. En matemática tienen dificultades 16 ítems de 70 aplicados, el 22.85%. En comunicación y lenguaje se eliminaron 2 ítems de 76 aplicados, el 2.63%. En ciencias naturales se eliminó 1 ítem de 65 aplicados, el 1.54%.

Referencias bibliográficas.

Alcántara, B. (2015). *Construcción del índice socioeconómico través del análisis factorial para la interpretación de resultados nacionales.* Guatemala: Ministerio de Educación, Digeduca.

Corral, Y. (2009). Validez y confiabilidad de los instrumentos de investigación para la recolección de datos. *Revista Ciencias de la Educación, 19*(33), 228-247. Recuperado el 2018, de http://servicio.bc.uc.edu.ve/educacion/revista/n33/art12.pdf

Dirección General de Evaluación e Investigación Educativa. (2019). *Informe Línea Base de Pruebas AVANZO, 2018. Versión final.* Guatemala: DIGEDUCA.

Dirección General de Investigación y Evaluación Educativa. (2014). *Infografía Evaluación de tercero básico 2013.* Guatemala: DIGEDUCA.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

ISSN: 2448-6574

Instituto Nacional de Estadística. (2011). *Mapas de pobreza rural.* Guatemala: INE. Recuperado el 2019, de https://www.ine.gob.gt/sistema/uploads/2015/09/28/V3KUhMhfgLJ81djtDdf6H2d7eNm0sWDD.pdf

Instituto Nacional para la Evaluación de la Educación. (2006). *Características de los Excale.* México: INEE.

Ministerio de Educación. (2016). *Anuario Estadístico 2015.* Guatemala: Ministerio de Educación. Recuperado el 2017, de http://www.mineduc.gob.gt/portal/index.asp

Ministerio de Educación. (s.f.). *Competencias Docentes.* Guatemala: Ministerio de Educación, Dirección General de Currículo.

Morales, A., Moreno, M., & Santos, J. (2015). *Análisis Rasch.* Guatemala: Dirección General de Evaluación e Investigación Educativa, Ministerio de Educación.

Pacheco–Villamil, J. (2007). LA EQUIPARACIÓN DE PUNTUACIONES EN PROCESOS DE COMPARACIÓN DE PRUEBAS DIFERENTES. *Avances en medición*, 153-156.

Pardo, C., Rocha, M., Avendaño, B., & Barrera, L. (2005). *Manual de procesamiento de datos y análisis de ítems.* Santiago de Chile: Oficina Regional de Educación para América Latina y el Caribe, UNESCO,Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), Segundo Estudio Regional Comparativo y Explicativo (SERCE).

Prieto, G., & Delgado, A. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100. Recuperado el 2017, de http://www.psicothema.com/pdf/1029.pdf

Quero Virla, M. (2010). Confiabilidad y coeficiente Alpha de Cronbach. *Telos, 12*(2), 248-252. Recuperado el 2018, de http://www.redalyc.org/pdf/993/99315569010.pdf

Universidad de San Carlos de Guatemala. (2009). *Rediseño curricular del Programa de Desarrollo Profesional del Recurso Humano del Ministerio de Educación -PDP-.* Guatemala: Universidad de San Carlos de Guatemala, Escuela de Formación de Profesores de Enseñanza Media.

Ureta, F., & Espinoza, E. (2015). *Revisión teórica internacional y nacional sobre la evaluación del desempeño docente y líneas generales de una propuesta de implementación en Guatemala.* Guatemala: Ministerio de Educación, Dirección General de Evaluación e Investigación Educativa.