Desarrollo de pruebas para evaluar resultados del desempeño de estudiantes costarricenses en el aprendizaje con tecnologías digitales

Melissa Villalobos García[1], Olmer Núñez Sosa[2], Gina Sequeira Guzmán[3] & Melania Brenes Monge[4]

1. Investigadora, Fundación Omar Dengo, San José, Costa Rica.
melissa.villalobos@fod.ac.cr
2. Investigador, Fundación Omar Dengo, San José, Costa Rica.
olmer.nunez@fod.ac.cr
3. Investigadora, Fundación Omar Dengo, San José, Costa Rica.
gina.sequeira@fod.ac.cr
4. Investigadora, Fundación Omar Dengo, San José, Costa Rica.
melania.brenes@fod.ac.cr

Recibido: 14 de octubre del 2017 **Corregido:** 20 de noviembre del 2017 **Aceptado:** 25 de enero del 2018

Resumen

El presente artículo realiza un aporte en el ámbito investigativo nacional sobre la construcción de pruebas específicas de evaluación formativa, y sobre algunos modelos de análisis psicométricos útiles para apoyar su diseño y ejecución en el ámbito de la educación. Se comparte la experiencia y las lecciones aprendidas del área de Investigación y Evaluación de la Fundación Omar Dengo con respecto a la creación de un instrumento para aproximar resultados de aprendizaje en estudiantes beneficiados del Programa Nacional de Informática Educativa (PRONIE MEP-FOD). Específicamente, se aborda la experiencia en el desarrollo y pilotaje de una prueba para estudiantes participantes de los Laboratorios de Informática Educativa en sexto grado, durante el año 2016, para lo cual se detallan los procedimientos desarrollados desde la construcción de los indicadores requeridos para aproximar el modelo conceptual, hasta la construcción de los ítems. Se concluye que este tipo de desarrollos requieren de altos criterios de rigurosidad y atención al detalle, con lo cual la documentación de los procedimientos llevados a cabo, así como de algunos de los criterios que se tomaron en cuenta en la toma de decisiones en cada fase, poseen valor tanto para la ejecución actual de las evaluaciones, como para procesos similares que se desarrollen a futuro, incluso en otros contextos.

Palabras clave. Evaluación del estudiante, Indicadores educativos, Tecnología de la información, Informática educativa, Método de medición, Psicometría.

Abstract

Development of tests to evaluate performance standards of Costa Rican students in digital technologies learning

This article aims to make a contribution to the national research field on the construction of specific tests of formative evaluation, and on some models of psychometric analysis useful to support their design and execution in the field of education. The experience and lessons learned from the

Research and Evaluation Unit of the Omar Dengo Foundation are shared with respect to the creation of an instrument to approximate learning outcomes in students benefited from the National Program for Educational Informatics (PRONIE MEP-FOD). Specifically, the experience in the development and piloting of a test for students of the Laboratories of Educational Informatics in sixth grade, during 2016 is addressed. The procedures developed from the construction of the indicators required to approximate the conceptual model, until the construction of the items are developed. It is concluded that this type of construction requires high criteria of rigurousity and attention to detail, with which the documentation of the procedures carried out, as well as some of the criteria that were taken into account in the decision making in each phase, have been valued both for the current execution of the evaluations, and for similar processes that are developed in the future in other contexts.

Key words. Student evaluation, Educational indicators, Information Technology, Computer Science Education, Measuring methods, Psychometrics.

INTRODUCCIÓN

Los recientes y acelerados cambios en las dinámicas socioeconómicas de la sociedad del conocimiento, donde las tecnologías de la información y la comunicación (TIC) juegan un rol fundamental, han dado lugar a importantes movimientos internacionales en el campo educativo (FOD, 2015). Los estudios realizados en el área se han encaminado a utilizar indicadores educativos y metodologías de evaluación "tradicionales" para medir el efecto de los programas de inclusión de TIC. Por ejemplo, aquellos relacionados con rendimiento académico, deserción y aumento del acceso digital (Computadoras para educar, 2014; CPA-Ferrere, 2010; Rivoir y Lamschetein, 2012).

El Programa Internacional para la Evaluación de Estudiantes (Organización para la Cooperación y el Desarrollo Económico [OCDE], 2014, 2013, 2006), con las pruebas PISA ha podido medir el dominio de procesos, el entendimiento de conceptos y habilidades en áreas como lectura, matemáticas y competencia científica, a través de pruebas de papel y lápiz (OCDE, 2006). No obstante, recientemente se han planteado evaluaciones que empiezan a derivar en hallazgos más específicos en relación con el desarrollo de habilidades para el uso de las TIC en el aprendizaje, partiendo de metodologías y enfoques menos tradicionales.

Sunkel y Trucco (2012) plantean que existe una corriente investigativa emergente que busca resultados en nuevos tipos de aprendizajes, por ejemplo las competencias del siglo XXI; dentro de las que se encuentran la resolución de problemas, pensamiento crítico, colaboración, entre otras. En esta línea, se han realizado estudios pioneros que, aunque no buscan medir solamente habilidades para el uso de las TIC en el aprendizaje, resultan ser un antecedente conceptual y metodológico esencial en investigaciones de esta naturaleza; como son los estudios del Consorcio para la "Evaluación y enseñanza de las destrezas del siglo XXI" (ATC21S, por sus siglas en inglés, http://www.atc21s.org/).

Desde los modelos de evaluación de competencias del siglo XXI se ha propuesto la construcción de pruebas y modelos de análisis psicométricos suficientemente sofisticados para dar cuenta de las habilidades mostradas (Griffin & Care, 2015; Bujanda & Campos, 2015; Griffin, Woods, Mountain & Scoular, 2013; Griffin, 2007; enGauge 21st Century Skills, por Ncrel, Metiri Group, 2003; el Partnership for 21st Century Skills, 2002). Lo anterior evidencia la existencia de especial atención de la investigación internacional por aproximar de manera más precisa los resultados de las intervenciones que se generan en la búsqueda del mejoramiento de la calidad educativa.

En Costa Rica, el Programa Nacional de Informática Educativa del Ministerio de Educación Pública y la Fundación Omar Dengo (en adelante PRONIE MEP-FOD) no ha sido la excepción, siendo uno de los

pilares de su gestión el desarrollo de procesos de investigación y evaluación para realimentar y mejorar constantemente sus propuestas educativas (Muñoz et al., 2014). Desde este programa se ha trabajado en la consolidación de las condiciones necesarias para llevar a cabo evaluaciones de resultados a gran escala que permitan dar cuenta del nivel de logro en las poblaciones meta.

El PRONIE MEP-FOD ha incorporado desde 1988 las TIC en los procesos de enseñanza y aprendizaje de los centros educativos públicos del país, con el fin de incrementar la calidad de la educación costarricense apoyando el desarrollo de capacidades mediante el uso de tecnología (Muñoz et al., 2014). La puesta en práctica de este modelo se ha logrado a partir de la atención de una serie de condiciones y la provisión de recursos a los centros educativos, en los que se han incluido computadoras como un objeto para construir y pensar (Zúñiga, 2001).

Al 30 de setiembre del año 2017, la cobertura del Programa corresponde a un 86,6% de la educación pública diurna desde preescolar hasta el III Ciclo, incluyendo Aula Edad, Aula Integrada y Educación Especial en Secundaria. Dentro de esa cobertura se contemplan los laboratorios de informática educativa (LIE).

La intervención educativa, en el contexto del LIE, ha estado apoyada en el aprendizaje basado en proyectos y la creación de productos digitales mediante la resolución de problemas con programación. En el año 2009, la propuesta se enriqueció con la publicación de los *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales* (Fundación Omar Dengo [FOD], 2009); los cuales "especifican qué se espera que éstos [*estudiantes*] sepan acerca de las tecnologías digitales y puedan hacer con ellas, para aprovecharlas en sus procesos de aprendizaje y continuar aprendiendo a lo largo de la vida" (FOD, 2009, p. 6). En síntesis, los estándares son los resultados de aprendizaje.

Los estándares de desempeño constituyen una serie de perfiles de salida para cada ciclo del sistema educativo costarricense, a saber: preescolar, I y II ciclo (primaria), y III ciclo y Educación Diversificada (secundaria). Se agrupan en tres dimensiones: resolución de problemas, productividad, ciudadanía y comunicación (ver síntesis de los estándares de desempeño en la Tabla 1).

TABLA 1
Síntesis de la conceptualización del modelo de Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales

Dimensión	Estándar
Productividad	**Estándar 1:** Se enfoca en el desarrollo de producciones digitales. **Estándar 7:** Orientado al uso de tecnologías digitales de manera responsable.
Resolución de problemas e investigación	**Estándar 2:** Centrado en el desarrollo de un proyecto utilizando tecnologías digitales. **Estándar 4:** Orientado al desarrollo de productos programados. **Estándar 5:** Se enfoca en la evaluación crítica de la información disponible en medios digitales.
Ciudadanía y comunicación	**Estándar 3:** Centrado en la comprensión de entornos colaborativos. **Estándar 6:** Se enfoca en la comprensión de repercusiones de la tecnología en la vida de las personas.

Fuente: Síntesis elaborada a partir de la conceptualización de FOD (2009).

La implementación de esta propuesta educativa ha sido una condición de viabilidad importante para aproximar los resultados de aprendizaje esperados de la integración de la tecnología en el contexto costarricense, ya que tras la implementación de la propuesta durante al menos cinco años, lo que prosiguió fue elaborar evaluaciones específicas para dar cuenta del nivel de logro de la población estudiantil.

En el año 2014, desde el Área de Investigación y Evaluación de la FOD, se inició un proceso sistemático de diseño y ejecución de evaluaciones formativas, por medio de pruebas especializadas, para aproximar los resultados de aprendizaje esperados en los estudiantes beneficiados de los LIE. Partiendo de este contexto, el objetivo del presente artículo es realizar un aporte a la gestión de conocimiento en el ámbito investigativo nacional sobre la construcción de este tipo de instrumentos específicos, y algunos modelos de análisis psicométricos útiles en el campo de la educación.

Se presentan los principales procedimientos llevados a cabo para desarrollar una prueba para evaluar habilidades en estudiantes de sexto grado durante el año 2016. De este modo, se detalla el procedimiento de construcción y pilotaje del instrumento, así como las principales lecciones aprendidas que se desprenden del proceso articulado que el equipo de investigadores de la Unidad de Evaluación de la FOD, bajo la dirección de Magaly Zúñiga Céspedes, ha estado desarrollando en los últimos años.

MÉTODOS

Diseño

Se desarrolló una prueba que colocara a la población estudiantil frente a una serie de retos para evidenciar sus conocimientos asociados a cada estándar de desempeño. Partiendo de un diseño de investigación mixto de tipo secuencial, en un primer momento se elaboraron insumos a partir de la recolección y análisis de datos cualitativos para, posteriormente, construir y aplicar un instrumento cuantitativo de mayor alcance (Hernández, Fernández y Baptista, 2010).

La construcción de un instrumento de esta naturaleza requiere altos niveles de rigurosidad en todas sus fases. En este caso, se tomó como base el enfoque de diseño centrado en la evidencia, cuya premisa es: "si el sujeto obtiene un resultado X, es porque conoce y puede hacer Y" (Mislevy, Almond & Lukas, 2003). En este diseño interactúan varios modelos, los cuales fueron considerados durante las distintas fases de la construcción de la prueba (Mislevy, Almond & Lukas, 2003; Zieky, 2014):

1. **Modelo de la población estudiantil.** Incluye variables relacionadas con los conocimientos, prácticas y disposiciones que dan evidencia de lo que se desea medir mediante la prueba.
2. **Modelo de la evidencia.** Contiene productos observables que pueden obtenerse tras una serie de tareas; es decir, son indicadores de logro basados en conocimientos, prácticas y disposiciones que permiten generar la evidencia necesaria para dar cuenta de los resultados.
3. **Modelo de las tareas.** Conforma la estructura de situaciones o retos que se presentarán a la población evaluada para obtener la evidencia de los indicadores de logro.
4. **Modelo de ensamblaje.** Contiene las especificaciones del instrumento, tales como instrucciones, tiempo de aplicación, formato y formas de respuesta, distribución de las preguntas, habilidades medidas, respuestas correctas y algunas especificaciones estadísticas requeridas.

Procedimiento

Definición del modelo conceptual

El insumo base para esta evaluación fue el documento *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales* (FOD, 2009), así como las *Guías Didácticas* (FOD, 2011). Como parte de la revisión conceptual, se realizaron mapas de habilidades para valorar la relación entre lo que se

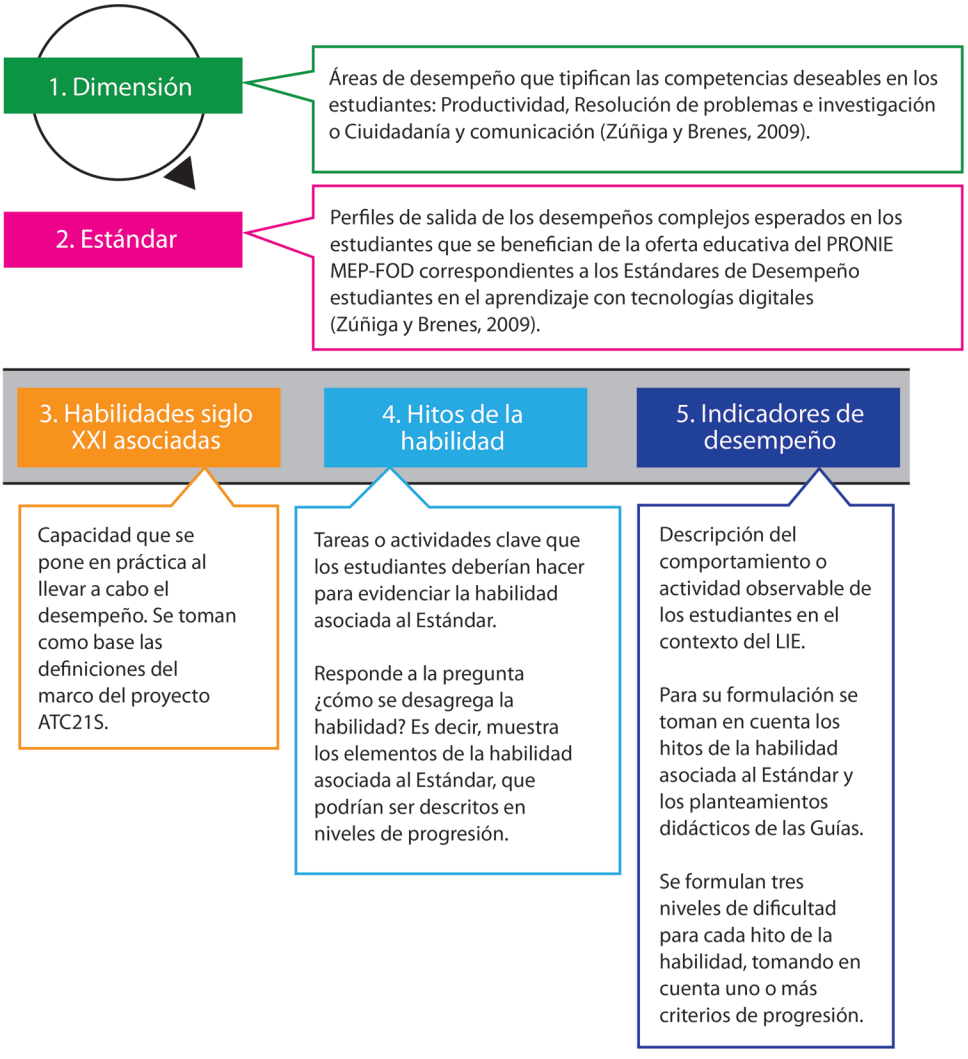


Figura 1. Detalle de la estructura conceptual para la evaluación. Fuente: Fundación Omar Dengo, 2017.

planteaba en cada estándar y las guías con otros modelos internacionales, específicamente con el *Marco de Habilidades de Siglo XXI* publicado por ATC21S (Griffin y Care, 2012).

El análisis de la interacción entre lo definido por los estándares de desempeño y el Marco de Habilidades del Siglo XXI permitió derivar de cada estándar una serie de actividades clave o hitos útiles para delimitar el objeto de investigación. Para cada uno de los hitos se construyó un conjunto de indicadores de evaluación como vía para aproximarlos y, posteriormente, cada uno de estos indicadores se transformaría en un ítem. Como se muestra en la Figura 1, el proceso permitió contar con una estructura conceptual clara y robusta.

Desarrollo de indicadores para aproximar resultados

Este paso involucró una fase de análisis y construcción de indicadores que dieran cuenta de los hitos, es decir, de las actividades clave observables de cada estándar de desempeño. En esta etapa, se realizó una revisión para depurar y alinear conceptualizaciones con los modelos más recientemente estudiados.

También se tomaron en cuenta las descripciones del modelo KSAVE (Knowledges, Skills, Attitudes, Values and Ethics –propuesto por Griffin & Care, 2015) como referente para su formulación.

En total se desarrollaron 29 hitos, y para cada uno de éstos se crearon tres indicadores de logro desarrollados en tres niveles: inicial, medio y esperado. De esta manera, se generaron 87 indicadores de desempeño, tomando en cuenta los siguientes criterios de progresión para definir el desempeño del estudiante:

- ***De lo individual a lo colaborativo.*** Los estudiantes van desde estrategias de trabajo individuales hasta procesos de aprendizaje en los que colaboran con otros para alcanzar una meta común.
- ***De lo declarativo a lo procedimental.*** Los estudiantes avanzan desde verbalizar sus conocimientos sobre cierto tema, hasta ponerlos en práctica cuando se enfrentan a diferentes situaciones de aprendizaje.
- ***De lo guiado por el docente a lo autónomo.*** Toma en cuenta desde el realizar tareas con apoyo de su docente hasta realizarlas sin requerir de esta mediación.
- ***De lo intuitivo a lo argumentativo.*** Avanzan desde buscar respuestas al azar, acierto y error, o incluso por exploración, hasta utilizar criterios argumentados al seguir un procedimiento concreto en la resolución de una tarea.
- ***De la simplicidad a la complejidad de la tarea.*** Los estudiantes van desde el realizar tareas sencillas hasta la resolución de tareas en las que debe integrar más elementos para buscar una respuesta.
- ***De lo apegado a normas a lo propositivo.*** Avanzan desde el seguir normas establecidas sin cuestionarlas, a proponer criterios propios que aportan valor a la tarea a la que se enfrentan.
- ***Didáctica de la programación.*** Los estudiantes avanzan desde el planteamiento de un problema o un plan para resolver una tarea, hasta la ejecución de una solución, mediante el uso de un lenguaje de programación.

Construcción del instrumento con criterios de validez y confiabilidad

La formulación de indicadores por niveles y con ciertos criterios de progresión, permitió establecer a priori el nivel de complejidad de los ítems (inicial, medio y esperado). La fase implicó crear al menos uno por cada indicador e involucró al equipo completo de investigadores de la Unidad de Evaluación. En total se generaron 87 ítems, a partir de criterios de calidad especificados en un manual desarrollado para este fin (FOD, 2016), dentro de los que destacan:

- ***Criterio de unidimensionalidad.*** El ítem debe medir únicamente un aspecto del modelo conceptual.
- ***Criterio de independencia.*** Un ítem no debe incidir ni sugerir la respuesta de otro.
- ***Criterio de selección única.*** Se debe hacer uso de las habilidades adquiridas mediante la participación en el Programa, para identificar la respuesta correcta.

Una vez elaborados, los ítems fueron sometidos a un proceso de validación para asegurar su calidad y validez. Mediante una estrategia de distribución por parejas de jueces, se dio un puntaje según la valoración de los criterios definidos en un baremo de validación (ver Figura 2). Posteriormente, se llevó a cabo un análisis de concordancia entre jueces (con procesos de conciliación en los casos necesarios) para revisar y corregir los ítems y obtener su versión final.

Fundación Omar Dengo

Baremo de validación de ítems
Objetivo: Validar el contenido de la prueba
Ítem:
Evalaudor(a):

Instrucciones: Califique cada uno de los ítems según los atributos que se le presentan.

Atributo	No cumple con el criterio	Tiene un bajo nivel	Tiene un nivel moderado	Tiene un alto	Observaciones
Claridad	1	2	3	4	
Formulación	1	2	3	4	
Distractores	1	2	3	4	
Pertinencia	1	2	3	4	
Unidimensionalidad	1	2	3	4	
Realidad	1	2	3	4	
Sensibilidad	1	2	3	4	
Impacto	1	2	3	4	

Puntaje:	Utilizar (Más de 28)	Revisar (De 16 a 28)	Descartar (menos de 16)
Nivel de complejidad asignado por el juez	Fácil	Moderado	Difícil

Figura 2. Baremo para calificación de ítems en análisis entre jueces. Fuente: Fundación Omar Dengo, 2016.

Cuando se tuvieron las versiones finales fueron distribuidas en tres formularios; con el propósito de respetar el tiempo disponible para la aplicación de la prueba (máximo de 80 minutos) y reducir el riesgo de que el factor cansancio incidiera en las respuestas de los evaluados.

Dado el interés de realizar una aplicación virtual, los formularios se realizaron en la plataforma en línea *Lime Survey* por las posibilidades que ésta brinda, como son la exportación de la base de datos al programa estadístico *Statistical Program for Social Sciences* (SPSS), la posibilidad de copiar ítems entre un formulario y otro, la identificación de usuarios, entre otros. No obstante, es importante aclarar que otras herramientas para generar formularios en línea también pueden ser funcionales para aplicaciones digitales de este tipo, con lo cual la toma de decisión sobre esto debe estar fundamentada en los requerimientos específicos que se definan para la prueba.

En cuanto a las estructura de los formularios, cada uno contaba con una sección común dedicada a recopilar información sociodemográfica y sobre acceso y uso de tecnología, información sobre las clases de informática educativa y sobre otras experiencias educativas relacionadas con tecnología.

La sección más robusta, correspondiente a los ítems de evaluación de los estándares de desempeño, fue la que se trabajó dividida en los tres formularios (pero con representación de cada estándar en cada uno de ellos) y es con la que se pretendía alimentar el banco. En la Figura 3 se puede observar la estructura de contenido de los formularios.

Es importante mencionar que en el apartado correspondiente a la evaluación de los estándares, se conservó en los tres formularios una base común de ítems denominados "ítems de anclaje", cuya función es permitir comparaciones mediante análisis de equiparación. Estos se seleccionaron según el comportamiento que tuvieron en el análisis entre jueces, tomando aquellos que presentaron mejores resultados. La estrategia implica asegurar no sólo la validez de cada formulario, sino también la viabilidad de poner a una misma escala el nivel de logro obtenido en los distintos formularios.

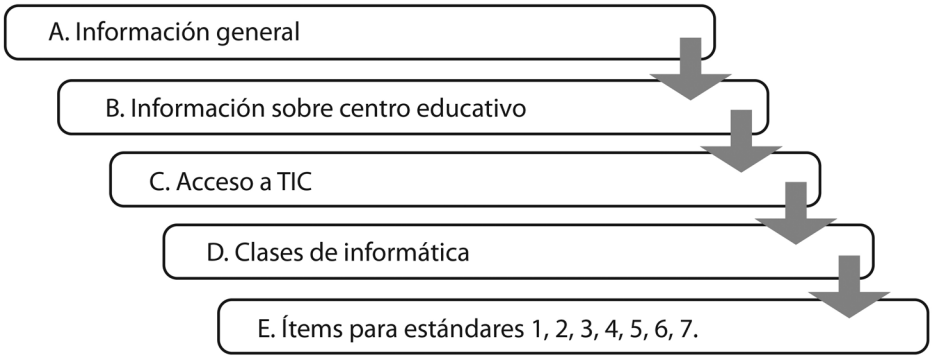


Figura 3. Apartados que conforman los formularios de la prueba. Fuente: Fundación Omar Dengo, 2017.

Población y muestra

Para el proceso de pilotaje se contó con una participación de 1619 estudiantes de 45 centros educativos, de los cuales 49,1% corresponde a mujeres (n=795). Las edades de los participantes se ubicaron entre los 11 y 13 años de edad (=11,9 años con D.E=0,7) y eran egresados del II ciclo (la prueba se aplicó al inicio del sétimo año para asegurar que hubieran completado el II ciclo). Además de la muestra de estudiantes, 33 docentes rellenaron un cuestionario sobre la ejecución de la propuesta de LIE y sobre el desempeño de sus estudiantes en estas lecciones.

RESULTADOS

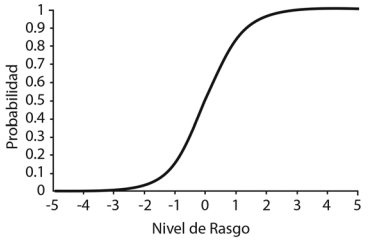
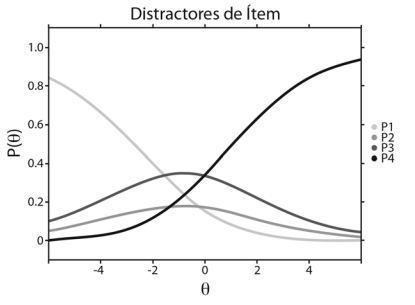
El proceso de pilotaje no buscó presentar resultados interpretativos del nivel de logro de la población estudiantil que respondió el instrumento, por lo que no se generaron aproximaciones de sus correspondientes niveles de habilidad; sino que el alcance se definió como un proceso de calibración de los ítems para enriquecer la conformación de un banco pertinente para procesos de evaluación posteriores.

Validación de la prueba

La calibración de ítems mediante los procedimientos de análisis de Teoría Clásica de los Test (TCT) y de Teoría de Respuesta al Ítem (TRI), permitió obtener información como el alfa de Cronbach de las dimensiones, el nivel de dificultad y el poder de discriminación de los ítems. Aunado al análisis de distractores y diferencial fue posible tomar decisiones sobre la pertinencia de incluirlos tal como estaban en el banco o valorar si deberían modificarse o excluirse. En la Tabla 2 se muestran los criterios empleados en análisis de calibración, así como algunas de las pautas establecidas para su interpretación.

Los análisis de calibración fueron complementados con la estrategia de laboratorios cognitivos, los cuales incluyeron observaciones y entrevistas cognitivas con estudiantes, para conocer sobre el entendimiento del vocabulario, las instrucciones y la ejecución de la prueba. Para las entrevistas cognitivas se utilizó como base la guía desarrollada por Smith y Molina (2011) y los datos recopilados se sistematizaron en un archivo de Excel para, posteriormente, realizar los ajustes correspondientes.

TABLA 2
Criterios para el análisis de calibración de los ítems

Sección	Subsección	Consideraciones para la interpretación
Teoría Clásica de los Test	% de Respuestas	El porcentaje de respuestas correctas debe estar acorde al nivel de dificultad esperado.
	Punto biserial	Debe ser mayor a 0,10.
	Alfa de Cronbach	El alfa de cada ítem no debe ser mayor que el alfa total de la prueba. Se recomiendan valores mayores de 0,8 (dependiendo del contexto, puntajes mayores de 0,7 pueden ser aceptables).
Modelo de 2PL (como modelo de TRI) Dificultad		Se esperan valores entre -2,5 y 2,5.
	Discriminación	En etapas iniciales de creación de un banco de ítems se recomiendan valores entre 0,5 y 2; mientras que en etapas más avanzadas entre 0,65 y 1,5.
Curvas de análisis de distractores	La curva para cada ítem	Para la opción correcta debe verse creciente hacia la derecha; indicando que aquellos estudiantes con mayor nivel de habilidad poseen mayor probabilidad de responderla. Mientras que las demás líneas (los distractores) muestran un buen comportamiento cuando decrecen gradualmente conforme aumente X, lo cual indica que aquellos estudiantes con menor nivel de habilidad poseen una mayor probabilidad de seleccionarlos.  
Curvas de información	Ver la curva de cada ítem	Lo ideal es que tenga una distribución normal, es decir, se esperan formas similares a la campana de Gauss.
Análisis diferencial	Para cada ítem	Revisar que no se esté beneficiando a ningún subgrupo de la población.

Fuente: Elaboración propia a partir del modelo de evaluación y calibración de los ítems.

Como resultado concreto del pilotaje se obtuvo que del total de 87 ítems creados, 54 se comportaron estadísticamente según lo esperado, es decir, cumplieron con todos los criterios de calidad y de calibración establecidos. En el caso de los restantes 33 fue necesario realizar un análisis de los elementos que mostraron algún tipo de problema para explicar su comportamiento diferente al esperado. Producto del proceso, se generaron algunos cambios o correcciones y, en los casos en los que se consideró oportuno, se tomó la decisión de crear un ítem nuevo.

En la Figura 4 se muestra un ejemplo de un ítem que presentó un comportamiento diferente al esperado en el análisis de calibración. En este caso, se revisaron los resultados del análisis de calibración y se encontró que mostró una discriminación negativa, es decir, que no discriminaba a los estudiantes con alta habilidad de aquellos con niveles más bajos.

Información sobre el ítem:

Estándar:	5	Dificultad: Medio	Habilidad: Pensamiento crítico y resolución de problemas
Hito clave:	4. Representan y formulan varias soluciones posibles para un problema.		
Indicador:	b. Identifican formas alternativas de responder al problema que se les propone.		
Criterios de progresión:	Didáctica de la programación De lo declarativo a lo procedimental		

Enunciado:

Marcela quiere que el gato camine al otro lado de la pantalla y toque la pieza de lego:


Se mueva caminando hasta la pieza de lego

Toque la pieza de lego

Para lograrlo, programó el siguiente bloque:

¿Cuál otro bloque de programación puede lograr que el gato haga lo mismo?

Opciones de respuesta:

A. (X)

B. ( )

C. ( )

D. ( )

Respuesta correcta:

El bloque A.
La opción A mueve el gato pocos pasos sin llegar al objeto, la opción B es un distractor, la opción D desliza el Gato sin caminar.

Figura 4. Ejemplo de ítem cuya calibración mostró comportamiento diferente al esperado.
Fuente: Formulario 1 de evaluación, Fundación Omar Dengo, 2017.

Por tanto, se concluyó que era necesario reformular las opciones de respuesta para aumentar el nivel de complejidad, ya que el análisis de distractores evidenció que estos parecen ser muy obvios para la población estudiantil. También se recomendó valorar la opción de generar una situación nueva para evaluar el hito, tomando en consideración que la situación propuesta inicialmente en el ítem parece resultar muy fácil para los estudiantes de esta edad.

El procedimiento de revisión descrito en el ejemplo anterior se llevó a cabo para los restantes 32 ítems, los cuales mostraron algún comportamiento atípico en la calibración. La toma de decisiones sobre los cambios y mejoras requeridas en cada uno contó con un proceso de revisión y aceptación por parte del equipo encargado de los procesos de construcción de ítems.

Por otra parte, el pilotaje permitió realizar ajustes para equilibrar los tres formularios en cuanto a la extensión y la valoración de contenidos realizadas en cada uno. Además, las salidas de datos y la coherencia entre los reportes de información hechos por estudiantes y docentes y los laboratorios cognitivos, permitieron identificar que los ítems son pertinentes en cuanto a claridad y comprensión para la población evaluada.

Validación del modelo conceptual

Del proceso de construcción y pilotaje de la prueba fue posible comprobar el modelo conceptual, es decir, se encontró la evidencia empírica necesaria para demostrar que cada ítem efectivamente estaba midiendo la dimensión para la cual habían sido construidos. La validación conceptual implicó valorar la configuración de los hitos completos, además de tomar algunas decisiones en los casos en los que se evidenciara algún problema de ajuste.

En este caso, se tomó la decisión de excluir dos hitos completos por problemas identificados en los tres ítems que los conformaban. Partiendo de lo anterior, la configuración final del modelo conceptual, por ende del banco, se redujo a 27 hitos (de los 29 planteados inicialmente), lo cual se traduce en 81 indicadores de evaluación con sus correspondientes 81 ítems. En la Figura 5 se muestra la configuración final del modelo conceptual.

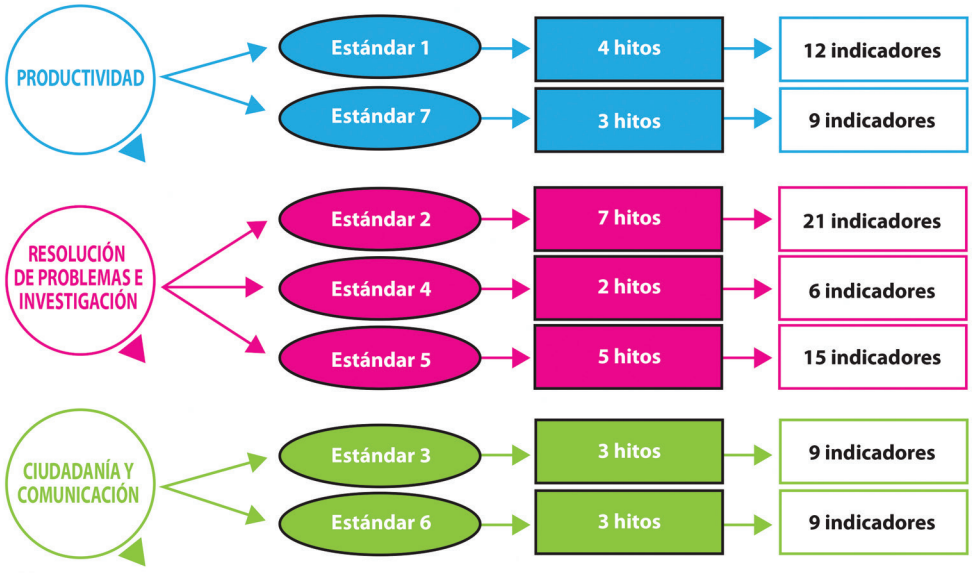


Figura 5. Configuración del modelo de evaluación de los estándares de desempeño en el 2016.
Fuente: Fundación Omar Dengo, 2017.

DISCUSIÓN DE LOS RESULTADOS

La alta cobertura del PRONIE MEP-FOD y la convicción de que indicadores aislados como rendimiento académico, o nivel de repitencia, no necesariamente dan cuenta del favorecimiento de habilidades específicas en el aprovechamiento de las TIC en la población estudiantil, fueron razones importantes que justificaron los esfuerzos por desarrollar pruebas como la descrita en el artículo. En términos generales, la experiencia en el desarrollo de este tipo de instrumentos ha permitido identificar procesos metodológicos cada vez más eficientes y oportunos para aproximar los resultados de los beneficiarios del programa.

Contar con altos niveles de rigurosidad metodológica en el desarrollo de una evaluación cuya naturaleza es compleja, es fundamental. Así, el proceso de pilotaje se considera un paso importante en cuanto a la calibración del banco de ítems, ya que permite no sólo valorar su comportamiento, sino también llevar a cabo un análisis estadístico pertinente para validar empíricamente el modelo conceptual; comprobando la consistencia de dicho empleo para la construcción de la prueba.

A nivel institucional, los resultados de procesos asociados a la construcción, pilotaje y aplicación de instrumentos de evaluación han permitido derivar una serie de lecciones aprendidas sobre las cuales se pretende enfocar la discusión. Así por ejemplo, desde una perspectiva metodológica, destaca la necesidad de definir la muestra tomando en consideración criterios como la zona; es decir, incluyendo centros educativos rurales y urbanos, la conexión a Internet de los centros educativos (al ser un cuestionario en línea) y su distribución por provincia.

El número de participantes en cada proceso de recolección de datos se estima a partir del alcance o el tipo de análisis que se quiere realizar, siendo usualmente las muestras de poco menos de 2000 estudiantes cuando se trata de procesos de pilotaje de pruebas. Las decisiones sobre los tamaños de la muestra, deben intentar mantener un balance entre la cantidad de participantes requeridos para llevar a cabo ciertos tipos de análisis estadísticos, y el no someter a un gran número de participantes a recolecciones de manera innecesaria.

Por otra parte, en la fase de montaje de los instrumentos un paso importante es la revisión interna que se hace de cada uno para examinar aspectos visuales y de formato, la fluidez o navegación entre pantallas, así como aspectos de digitación y presentación (por ejemplo, ortografía, omisión de alguna palabra, activación de todas las opciones de respuesta, entre otros). Al respecto de esto, las tablas de especificaciones resultan particularmente útiles para garantizar que cada formulario contenga los aspectos requeridos según el alcance definido.

Se ha identificado que contar con más de un formulario es necesario cuando el proceso de calibración incluye una amplia cantidad de ítems. Esto es oportuno para no saturar a la población evaluada con una prueba excesivamente larga, lo cual podría además poner en riesgo la validez de las respuestas emitidas. Sin embargo, se debe considerar que el procedimiento de segmentación implica posteriormente llevar a cabo varias estrategias de análisis que permitan comprobar aspectos de validación en cada formulario, así como una adecuada equiparación para dar cuenta de los resultados de manera global.

Al respecto de las recolecciones de datos, tanto en los procesos de pilotaje como en las recolecciones formales para dar cuenta de resultados, se deben tomar en cuenta controles que permitan garantizar la calidad de la información. Uno de ellos es asegurar que cada participante reciba las mismas instrucciones, para lo cual se recomienda realizar videos para estudiantes, tutoriales para docentes y otros materiales útiles para garantizar que la población reciba las indicaciones con la misma calidad.

Otro control que contribuye a garantizar la calidad de la información es la rigurosidad con la que se lleva a cabo el procesamiento o sistematización de los datos obtenidos. En este sentido, se debe considerar

que en los procesos de pilotaje este procedimiento se presenta en dos vías: una más automática cuando cada participante completa el formulario en línea y esa información se almacena en una base de datos; y otra manual, que implica la sistematización de la información obtenida en las observaciones y las entrevistas cognitivas en plantillas específicas que permiten almacenar la información, por ejemplo, en bases de datos en formato de hojas de cálculo. La expectativa de todos estos registros es que la información sea útil al momento de realizar los análisis de cada ítem, con lo cual la claridad y el detalle se consideran características pertinentes.

Finalmente, se debe mencionar que en la medida en la que se alimente y calibren los bancos de ítems se pueden desarrollar instrumentos de evaluación con mayor estabilidad, validez y consistencia. En el contexto PRONIE MEP-FOD, esto resulta altamente importante porque contar con un banco calibrado y confiable es fundamental para valorar la intervención que se realiza por medio de los LIE en las poblaciones beneficiarias. Partiendo de esto, resulta oportuno destacar la necesidad de mantener los procesos de calibración como una práctica periódica vinculada a los procesos de evaluación de resultados en estudiantes, con la finalidad de contar con un banco de ítems cada más estable y confiable.

CONCLUSIONES

El objetivo del proceso de construcción de pruebas específicas para aproximar el logro de los estándares de desempeño de los estudiantes beneficiarios del PRONIE MEP-FOD se orienta a valorar sus aprendizajes en relación con las tres dimensiones definidas. En este sentido, el proceso de construcción teórico y metodológico descrito en este artículo fue guía para diseñar un conjunto de ítems que permitieran, de un modo pertinente, aproximar el modelo conceptual.

La validación empírica del modelo conceptual se considera un resultado importante del pilotaje, ya que para procesos de evaluación posteriores será posible emitir conclusiones fiables sobre el nivel de desempeño estudiantil en cada una de las dimensiones estudiadas.

La experiencia en el desarrollo y la aplicación de este tipo procesos que implican cierto nivel de dificultad ha dado importante evidencia de que el éxito depende en gran medida de la gestión de un trabajo colaborativo que implica requerimientos y coordinaciones logísticas específicas, tanto dentro de la organización como con los centros educativos. En este sentido, el orden y la atención al detalle son fundamentales para lograr que cada procedimiento se realice de la forma esperada y se recupere la información según los criterios de calidad establecidos.

La documentación de estos procesos de evaluación y las lecciones aprendidas que se generan a partir de la experiencia pretenden funcionar como un control ante la complejidad de la tarea. Así, se busca que la experiencia en el refinamiento y depuración de los modelos conceptuales y de los procedimientos metodológicos no sólo sean útiles para los procesos actuales de aproximación de logro, sino para el planteamiento de otros procesos similares a futuro.

Finalmente, es oportuno mencionar que parte del éxito de los procesos de recolección de información depende de la medida en la que los docentes y directores de los centros educativos se comprometan con el cumplimiento de los plazos y los controles de calidad. En este sentido, se externa un agradecimiento a los participantes del PRONIE MEP-FOD, quienes año con año han colaborado con las diferentes fases de recolección de información vinculadas a estos procesos evaluativos.

REFERENCIAS

Bujanda, M. & Campos, E. (2015). The adaptation and contextualization of ATC21STM by Costa Rica. In P. Griffin & E. Care (eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp.245-256). Dordrecht, Netherlands: Springer.

Computadores para educar. (2014). *Informe de gestión año 2013 Colombia*. Recuperado de http://www.computadoresparaeducar.gov.co/PaginaWeb/images/descargables/Informe_gestion_2013%20.pdf

CPA-Ferrere (2010). *Principales lineamientos estratégicos: Informe final Uruguay*. Recuperado de http://www.ceibal.edu.uy/Documents/Informe%20Plan%20Estrategico%20CEIBAL.pdf

Fundación Omar Dengo [FOD]. (2009). *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales*. San José, Costa Rica: Fundación Omar Dengo.

Fundación Omar Dengo [FOD]. (2011). *Guías didácticas para el trabajo en el laboratorio de Informática Educativa del PRONIE MEP-FOD*. San José, Costa Rica: Fundación Omar Dengo.

Fundación Omar Dengo [FOD]. (2015). *Evaluación de los estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales, PRONIE MEP-FOD: Marco de Referencia.* (Documento interno). San José, Costa Rica: Fundación Omar Dengo.

Fundación Omar Dengo [FOD]. (2016). *Manual para la creación de ítems.* Unidad de Evaluación, Fundación Omar Dengo (Documento interno). San José, Costa Rica: Fundación Omar Dengo.

Fundación Omar Dengo [FOD]. (2017). Nivel de logro de los Estándares de Desempeño en estudiantes de sexto grado del PRONIE MEP-FOD 2016: Informe de resultados. Fundación Omar Dengo (Documento interno). San José, Costa Rica: Fundación Omar Dengo.

Griffin, P. & Care, E. (2015). Assessment and Teaching of 21st Century Skills: Methods and Approach. Dordrecht, Netherlands: Springer.

Griffin, P. & Care, E. (2015). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht, Netherlands: Springer.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation. 33(*1), 87-99. DOI: 10.1016/j.stueduc.2007.01.007.

Griffin, P., & Care, E. (2012). *Assessment and Teaching of 21st Century Skills.* Dordrecht, Netherlands: Springer.

Griffin, P., Woods, K., Mountain, R. & Scoular, C. (2013) Module 1: *Using a Developmental Model to Assess Student Learning.* Recuperado de http://www.ATC21s.org/uploads/3/7/0/0/37007163/pd_module_1_for_web_2014.pdf

Hernández, R., Fernández, C. y Baptista, P. (2010). *Metodología de la Investigación* (5ta. edición). México D.F, México: Mc Graw Hill.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief Introduction to evidence-centered design* (ETS Research Report 03-16). Princeton, Estados Unidos: Educational Testing Service.

Muñoz, L., Brenes, M., Bujanda, M., Mora, M., Núñez, O., & Zúñiga, M. (2014). *Las políticas TIC en los sistemas educativos de América Latina: Caso Costa Rica.* Buenos Aires, Argentina: UNICEF.

Organización para la Cooperación y el Desarrollo Económico. [OCDE] (2006). *Programa para la Evaluación Internacional de Alumnos PISA: Marco de la evaluación. Conocimientos y habilidades en Ciencias, Matemáticas y Lectura.* Recuperado de http://www.oecd.org/pisa/39732471.pdf

Rivoir, A. L., y Lamschetein, S. (2012). *Cinco años del Plan Ceibal: algo más que una computadora para cada niño*. Recuperado de https://www.unicef.org/uruguay/spanish/ceibal-web.pdf

Smith, V. y Molina, M. (2011). *La entrevista cognitiva: guía para su aplicación en la evaluación y mejoramiento de instrumentos de papel y lápiz (Cuadernos metodológicos No. 5)*. Instituto de Investigaciones Psicológicas-UCR: San José, Costa Rica.

Sunkel, G. y Trucco, D. (2012). *Las tecnologías digitales frente a los desafíos de una educación inclusiva en América Latina: Algunos casos de buenas prácticas*. Santiago, Chile: Comisión Económica para América Latina y el Caribe (CEPAL).

Zieky, M. J. (2014). An Introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*(1), 79-87. DOI: 10.1016/j.pse.2014.11.003

Zúñiga, M. (2001). *Del construccionismo al constructivismo*. San José, Costa Rica: Fundación Omar Dengo.