Ciencia y Educación
Vol. 1, No. 1, septiembre-diciembre, 2017

Construyendo Sistemas de Monitoreo y Evaluación Eficaces para Proyectos de Educación: Experiencias del Proyecto Leer

Building effective monitoring and evaluation systems from education projects: Experiences from Project Read

Aída Mencía-Ripley[1]
Carlos Ruíz Matuk[2]
Laura V. Sánchez-Vincitore[3]

Recibido 18/10/2017 • Aprobado: 31/10/2017

Resumen

Los proyectos locales en materia educativa son cada vez más numerosos y ambiciosos en su alcance. Dichos proyectos presentan oportunidades idóneas para diseñar intervenciones efectivas, sistematizarlas y producir un alto volumen de trabajos científicos. Para ello, es crucial crear sistemas robustos de medición que permitan la toma de decisiones basadas en evidencia, la corrección de aspectos de la intervención que no se comportan de la manera esperada, y la medición misma del impacto del proyecto. Este abordaje permite realizar proyectos educativos efectivos, apegados a la ética y la protección de poblaciones vulnerables.

Palabras clave: cooperación educacional; psicometría; evaluación.

Abstract

Local education projects are increasing in number and scope. These projects are ideal spaces in which to design effective interventions, standardize them, and produce a high volume of scientific literature. To achieve this, it is imperative to create robust measurement systems. Said systems must allow the project to make evidence based decisions, correct intervention aspects that do not work as expected, and to properly measure impact. This approach allows implementers to implement effective projects ethically, with the protection and well-being of vulnerable populations in mind.

Keywords: educational cooperation; psychometrics; evaluation.

1. Universidad Iberoamericana (UNIBE), Santo Domingo, República Dominicana. ORCID: 0000-0001-7510-4072. a.mencia@unibe.edu.do

2. Universidad Iberoamericana (UNIBE), Santo Domingo, República Dominicana. ORCID: 0000-0003-2681-4953. c.ruiz4@unibe.edu.do

3. Universidad Iberoamericana (UNIBE), Santo Domingo, República Dominicana. ORCID:0000-0002-6343-1217. l.sanchez1@prof.unibe.edu.do

La realización del presente documento es posible gracias al generoso apoyo del pueblo americano, a través de la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID). Los contenidos son responsabilidad de UNIBE y no necesariamente reflejan la visión de USAID o del gobierno de los Estados Unidos.

Construyendo Sistemas de Monitoreo y Evaluación Eficaces para Proyectos de Educación: Experiencias del Proyecto Leer

En la República Dominicana se realizan numerosos proyectos educativos, aunque, en los últimos años, se ha visto un incremento en programas de investigación en el sistema de educación superior nacional, la creación de institutos dedicados exclusivamente a la investigación, y reformas estatales de gran alcance en el sistema educativo de nuestro país. Se le suma a esto, lo que la cooperación internacional invierte en la educación nacional. Todas estas variables crean un territorio fértil para múltiples intervenciones y acciones que generan información valiosa para el diseño e implementación de estrategias innovadoras, así como la toma de decisiones basadas en evidencia.

No poder sistematizar y evaluar el impacto de estas intervenciones de manera adecuada, constituye una pérdida importante al sistema educativo y para los científicos nacionales. Este escenario impediría la sistematización de buenas prácticas, la réplica de intervenciones eficaces, así como el estudio científico de los problemas educativos locales y sus posibles soluciones.

Las oportunidades de innovación y las posibles contribuciones de científicos dominicanos a la educación y la psicología, a partir de estas intervenciones, carecen de impacto, ya que los criterios de rigor científico para comunicaciones y publicaciones internacionales, se dificulta en terrenos educativos en países en vías de desarrollo. Esto, por la carencia de un sistema nacional de publicación científica, así como las pocas capacidades instaladas en los diversos sectores que participan de la educación. A esto se añade la ausencia de instrumentos locales con adecuada validación y fiabilidad, que permitan evaluar de manera consistente las variables de interés para investigadores, psicólogos y educadores.

Conociendo esta realidad, el Proyecto Leer ha creado un sistema de monitoreo y evaluación que permite la sistematización adecuada de los conocimientos generados por el proyecto, de forma que se pueda:

- Corregir aspectos de la intervención que no estén funcionando.
- Diseñar investigaciones de calidad que permitan generar datos fiables.
- Tomar decisiones programáticas partiendo de la evidencia.
- Crear una plataforma de divulgación de información a actores locales del sistema educativo.
- Producir investigaciones de alta calidad que permitan producir conocimiento y compartirlos con la comunidad científica nacional e internacional.
- Generar datos estandarizados que permiten estudios comparativos válidos.

Para lograr estos objetivos, se identificaron las investigaciones, la psicometría y un sistema de monitoreo global, como estrategias claves para lograr producir y compartir datos de calidad. Este tipo de sistema se apega a la ética de la investigación en diversos aspectos, uno de ellos es, la toma de decisiones basadas en la ciencia, que asegura que personas vulnerables (y que suelen ser el foco de intervención en proyectos de cooperación), no se expongan a dichas intervenciones sin beneficios probados. Asimismo, pone el bienestar de los mismos como el criterio de mayor importancia del quehacer del proyecto, y de la propia actividad investigativa (Morra Imas & Rist, 2009).

Este artículo propone describir el Proyecto Leer. Se discuten las características del mismo, en especial, su sistema de monitoreo y evaluación y sus componentes fundamentales, como la psicometría y la investigación. Por último, planteamos una discusión sobre el alcance e impacto de dicho sistema en la ejecución eficaz del proyecto, así como sus aportes globales a la generación de información científica local.

Descripción del Proyecto Leer

El Proyecto Leer es una iniciativa de la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID, por sus siglas en inglés). El objetivo del proyecto es mejorar la lectoescritura en estudiantes de primaria (grados 1 a 6), en escuelas públicas dominicanas ubicadas en el corredor Duarte.

El corredor Duarte es una designación geográfica que utiliza USAID en su demarcación de lugares

nacionales, en los cuales, focaliza el trabajo de cooperación. El área corresponde a las zonas aledañas a la Autopista Duarte. Actualmente, 396 escuelas participan en el proyecto, perteneciendo estas, a las regionales de San Cristóbal, La Vega, Santiago, Mao, Santo Domingo, Puerto Plata y Cotuí. De estas escuelas, 155 son de jornada extendida y 373 tienen formación preescolar.

El proyecto cuenta además con un componente de reducción de violencia de género, inclusión de niños y niñas con discapacidad y prevención de la violencia escolar, incluyendo el *bullying;* (género, inclusión, y escuela segura: GISS, por sus siglas en inglés). Finalmente, un componente comunitario implementado por un socio del proyecto (Visión Mundial), proporciona espacios comunitarios de apoyo a la lectura y fomenta una participación más activa de las familias y comunidades en la educación primaria.

Investigación

Para poder medir el impacto de la intervención, se realizó un estudio de línea base que contó con un diseño experimental de grupo control, el cual fue medido antes de la intervención, y será medido nuevamente al finalizar la misma. Este diseño, no obstante, solo permite evaluar el impacto cada dos años y medio, por lo que la toma de decisiones para corregir aspectos del proyecto que requieran pronta atención, se verían limitadas. Para subsanar esta brecha, se han diseñado instrumentos de monitoreo estandarizados que se aplican en cada sesión de acompañamiento con los docentes, los cuales, son evaluados todos los años.

Adicional a esto, cada indicador del proyecto tiene un proceso estandarizado que permite el levantamiento de datos anualmente, con un único modelo de cálculo, plasmado en la memoria del proyecto, para evitar cambios metodológicos que pudiesen limitar la comparabilidad de los datos. Finalmente, el equipo de monitoreo y evaluación ha creado una rúbrica para guiar de manera estandarizada cuáles productos de investigación se deben desarrollar, y para cuáles audiencias.

En primer lugar, se elaboran comunicaciones no científicas para actores clave fuera del sistema educativo. Esto permite dar a conocer los hallazgos generales del proyecto, así como informar a padres y madres sobre el progreso de la intervención. También se desarrollan comunicaciones técnicas para tomadores de decisiones, las cuales se presentan a través de informes específicos, altamente detallados o en eventos profesionales. Finalmente, cuando el proyecto genera una contribución sustancial al conocimiento, se procede a elaborar comunicaciones científicas en forma de presentaciones escritas en congresos, así como publicaciones para la comunidad científica nacional e internacional. Estas publicaciones se realizan a partir de análisis primarios, así como análisis secundarios.

Dentro del equipo de monitoreo y evaluación se creó la posición de "coordinador de investigación", cuya responsabilidad es integrar la literatura académica a los resultados que se obtienen, luego de cada levantamiento de datos. Esta integración es una fuente generadora de investigaciones asociadas al proyecto, que permite adelantarse y prevenir posibles amenazas a la validez interna y externa de las diversas mediciones que se realizan. En ese sentido, la estrategia ha identificado varias líneas de investigación de interés nacional con relación a la adquisición de la lectoescritura, que va desde el impacto de factores sociales y ambientales a la escolarización, o la función de la primera infancia, hasta el diseño de materiales apropiados al nivel lector, y estrategias de enseñanza específicas de la lectoescritura.

Luego de dos años de intervención, se han generado las siguientes comunicaciones presentadas en dos congresos internacionales:

- Desarrollo de textos decodificables
- Presentación de resultados de línea base
- Presentación sobre el modelo simple de la lectura
- Hallazgos del sistema de monitoreo y evaluación

Una quinta ponencia está en espera de presentación en un tercer congreso internacional. De igual manera, se han realizado presentaciones en foros y coloquios a nivel nacional.

El proyecto ha generado los siguientes reportes:

- Resultados de línea base GISS (infográfico)

- Reporte de línea base del componente comunitario
- Reporte de línea base del componente GISS (género, inclusión, y escuela segura)
- Resultados de línea base de lectura (infográfico)
- Informe al Ministerio de Educación: Línea Base
- Reporte de línea base del proyecto USAID-Leer
- Evaluación de los sesgos en los libros de texto de Lengua Española, de 1ero a 3er grado
- Reporte de fin de segundo curso (2016)
- Reporte de fin de segundo curso (2017)
- Infografía de reporte de fin de segundo (2016)
- Infografía de reporte de fin de segundo (2017)
- Informe de actividades internacionales con niños de primaria en República Dominicana y México

Dada la robustez del sistema de monitoreo y evaluación, contamos con datos de calidad para elaborar y enviar los siguientes trabajos a revistas académicas:

- Estudio sobre el modelo simple de la lectura utilizando modelado de ecuaciones estructuradas
- Estudio piloto sobre los libros decodificables
- Estudio sobre calidad de gestión y habilidad lectora
- Estudio sobre estatus socioeconómico y habilidad lectora. Estudio cualitativo sobre el dominio cultural "calidad de la gestión"

Psicometría y Validez de la Información.

Para poder generar información de calidad, es necesario contar con instrumentos que cuentan con las propiedades métricas adecuadas, para generar datos válidos y fiables. Dichos instrumentos deben tener un lenguaje que permita ser fácilmente entendido, y medir constructos que se definan por la cultura en la cual se van a medir.

En el contexto de proyectos educativos de gran alcance, como lo es el proyecto Leer, los propósitos para estas evaluaciones pueden contemplar detección, diagnóstico, intervención, evaluación, selección y certificación (Braden, 2003). En este proyecto, hasta ahora se han incluido evaluaciones dirigidas a la detección y el diagnóstico (línea base).

Para las evaluaciones puede recurrirse a métodos cualitativos o cuantitativos, además de procedimientos mixtos. Dentro de las alternativas para el diseño de las evaluaciones se encuentran: a) el desarrollo y validación de los instrumentos a usar; b) la selección de los mismos respondiendo a los propósitos planteados.

Sin importar cuál es la alternativa elegida, es importante tomar en cuenta varios criterios, especialmente atendiendo al problema de la proliferación de escalas e información actual, que hacen difícil al investigador determinar y localizar las mejores escalas para los constructos de interés, al tiempo que están dispersos en una multitud de revistas y libros, o presentaciones comerciales por diversos editores (Boyle, Saklofske, & Matthews, 2014).

Estos criterios generalmente incluyen tres grupos o categorías (Robinson, Shaver, & Wrightsman, 1991):

- Criterios relacionados a la construcción de los ítems (muestreo del contenido relevante, redacción de los ítems y análisis de los ítems).
- Criterios relacionados al control de las distorsiones motivacionales (aquiescencia y la deseabilidad social).
- Criterios relacionados a la métrica de los instrumentos (confiabilidad, validez, normas y muestreo probabilístico).

Con respecto a la construcción de los ítems, tomamos en cuenta el muestreo de contenido o las diferencias en las formas en que se redactan los reactivos o preguntas (Cohen & Swerdlik, 2006). En este sentido, se debe revisar si los creadores de las escalas a considerar, han presentado en sus trabajos de validación, la delimitación conceptual del constructo a evaluar. Por ejemplo, si en su publicación sobre el instrumento se describe intento previo de conceptualización del constructo de interés, si se establecen las relaciones esperadas entre el constructo medido por el test y otras variables, etc.

En cuanto a la redacción de los ítems, debe destacarse, que ha sido considerada como una de las tareas

más complejas en la elaboración de un instrumento (Díaz Esteve, 1997). Para una buena selección de un instrumento, debe tomarse en cuenta si se explican las razones por las cuales se adoptó un determinado formato y escala de respuesta para los ítems, si se ha evaluado la claridad de los enunciados de los ítems y su funcionamiento.

Un aspecto a tomar en cuenta es la traducción de las escalas correspondientes al estudio. En estudios interculturales, por ejemplo, el uso de instrumentos previamente desarrollados puede ahorrar mucho tiempo y esfuerzo. Sin embargo, estos instrumentos deben ser culturalmente aceptables y traducidos apropiadamente para ser válidos. Generalmente, una traducción directa de una escala, no garantiza la equivalencia de contenido de la escala traducida. Una estrategia muy utilizada se denomina traducción inversa, retro traducción o "back translation" (Brislin, 1970).

El modelo de Brislin, de traducción y retro traducción de instrumentos, es un método bien conocido (Brislin, 1970; Jones, Lee, Phillips, Zhang, & Jaceldo, 2001; Sousa & Rojjanasrirat, 2011). De acuerdo con este modelo, un experto bilingüe traduce el instrumento de la lengua de origen (SL) a la lengua de destino (TL), y un segundo experto bilingüe (sin acceso a la versión original) traduce a la lengua de origen. Si se encuentra un error en el significado en la versión retro traducida en comparación con el original, los términos que están en cuestión son retraducidos y vueltos a ciegas, para ser traducidos por otro experto bilingüe. Este proceso se repite hasta que no se encuentre ningún error en el significado.

Es conveniente aclarar que, aunque la retro traducción es un proceso cuidadoso, es posible que aparezca una equivalencia aparente entre la fuente y las versiones retro traducidas de un instrumento sin equivalencia real. Por ejemplo, el primer traductor y el traductor posterior pueden usar el mismo dialecto del lenguaje y tener interpretaciones regionales similares del significado de las palabras, o un buen traductor posterior puede dar sentido a una versión de idioma objetivo deficiente (Jones et al., 2001; Sperber, 2004). La relevancia de elegir traductores calificados para el proceso, es notoria.

En el proceso de elaboración del instrumento, una vez construidos los ítems y revisados por una comisión de expertos, estos deben ser puestos a prueba, aplicándolos a una muestra piloto de sujetos, para determinar su nivel de dificultad, así como su capacidad para discriminar entre los sujetos con mayor y menor aptitud, o presencia del atributo a medir. Esta estrategia ha sido empleada en todos los levantamientos de datos del Proyecto Leer. Es decir, se estudia sus características psicométricas, de modo que puedan seleccionarse los mejores ítems para formar el instrumento. Este paso del proceso, llamado también "aplicación experimental" o "Prueba piloto", debe cubrir dos fases (Díaz Estéve, 1997):

a. La aplicación de los ítems a una muestra piloto

Al seleccionar la muestra de participantes deben definirse las edades, los cursos y otras variables, para que se pueda tener cierta seguridad de que la muestra presenta las mismas características que la población. Al seleccionarla, hay que cuidar que la misma sea elegida de varios centros o localidades, para evitar que los resultados estén distorsionados por las condiciones específicas de un solo centro o de una sola localidad. El número de sujetos a elegir para formar parte de la muestra, debe ser lo suficientemente grande como para que las inferencias hechas sobre sus resultados sean admisibles. Cuando la prueba se va aplicar a grandes grupos se considera adecuado aplicarla a una muestra compuesta de cien a doscientos sujetos, procurando que sea más o menos estratificada en los valores de las variables controladas: edad, curso o categoría, de forma que los parámetros obtenidos resulten útiles para describir los ítems.

b. Análisis formal de los ítems

El análisis formal de los ítems busca hallar unos índices numéricos que permitan hacer una selección objetiva de ítems, y en cierto modo, segura desde el punto de vista de la metodología científica. El análisis formal y la selección de ítems se hace de acuerdo al marco teórico o sistematización teórica en que se sitúa el constructor: la Teoría Clásica de los Test (TCT), la Evaluación Referida al Criterio (ERC), la Teoría Clásica a los Ítems (TRI) y la Psicometría Cognitiva. En

todas ellas se calculan índices o parámetros relativos a la dificultad, a la capacidad discriminativa (interna o externa) y al análisis de las alternativas.

Cualquier informe sobre la elaboración de un instrumento debe cubrir entonces esas dos fases. Es decir, que además de que es importante emplear una muestra piloto, cuyas características y procedimientos a seguir, deben describirse claramente para obtener a través de ella esas cualidades de los ítems; asimismo, debe justificarse las decisiones sobre los ítems a mantener o eliminar. Para esto, se requiere realizar análisis estadísticos para evaluar el funcionamiento de los ítems con la muestra piloto, ofreciendo datos estadísticos del nivel, de acuerdo a las valoraciones de contenido de los jueces expertos.

Atendiendo a los elementos relacionados al control de la aquiescencia y deseabilidad social (Cronbach, 1950), se han ofrecido diferentes estrategias (Nederhof, 1985; Robinson et al., 1991). La aquiescencia se refiere a la predisposición a obedecer a una categoría de respuesta de acuerdo a la situación o contenido del ítem. Por otro lado, la deseabilidad social se refiere a la idea de que las personas tratan de hacer una buena impresión a los demás (Edwards, 1957).

Hace muchos años existió una controversia sobre la influencia de estos factores en la validez del instrumento (Dicken, 1963; Greenwald & Satow, 1970; Jackson & Messick, 1962). Sin embargo, recientemente se puede apreciar una aceptación de sus efectos sobre la estructura factorial de los autoreportes (Navarro-González, Lorenzo-Seva & Vigil-Colet, 2016).

Las estrategias para controlar la aquiescencia incluyen el intercambiar el orden de las opciones de las respuestas, incluir ítems reversos y la construcción de ítems de selección forzada. En el caso de la deseabilidad social, los controles incluyen también el uso de ítems de selección forzada, en las cuales, las alternativas han sido equiparadas en base a las estimaciones de deseabilidad social (estudiadas anteriormente).

Por último, un grupo de criterios muy importante para seleccionar un instrumento, están ligados a las métricas del mismo. En primer lugar, el instrumento debe tener consistencia o estabilidad. A esto se le denomina fiabilidad. La fiabilidad está ligada a la ausencia de error en la medida.

Existen diversas formas para saber si una prueba es confiable:

- Test-Retest: Es la administración de una prueba por segunda vez a los mismos participantes, para observar si los resultados son estables.
- Formas paralelas: Consiste en crear la misma prueba con un diferente muestreo de contenido, y comparar sus resultados.
- Dos mitades: Consiste en dividir la prueba en dos mitades (por ejemplo, los reactivos pares versus los impares), y obtener la correlación entre los ítems contiguos.
- Kuder-Richardson: Se emplea en pruebas cuyas preguntas sólo tengan dos respuestas o ítems dicotómicos (Verdadero y Falso; Correcto o Incorrecto).
- Confiabilidad entre calificadores: Cuando dos o más jueces califican las respuestas de diferentes personas. Hay varias formas de estimar la confiabilidad intercalificadores, interobservadores o interjueces, como se le ha llamado. Uno consiste en registrar el porcentaje de acuerdos entre los observadores. Sin embargo, este método tiene el inconveniente de que no toma en cuenta el acuerdo que ya se produciría al azar. Otro método consiste en la estadística Kappa, para situaciones donde dos observadores tienen que clasificar (Cohen, 1960).
- Alfa de Cronbach (Cronbach, 1951): Se emplea en escalas donde existen más de dos opciones de respuestas. Se puede decir que las fórmulas de Kuder-Richardson son versiones del coeficiente alfa (α) más general (Aiken, 2003).

La mayoría de los cálculos de la confiabilidad son coeficientes de correlación o afines. Las puntuaciones suelen variar desde 0 a 1, siendo más altos los indicadores de mayor confiabilidad. Se recomienda que las pruebas presenten una confiabilidad entre .7 y .9, aunque puede ser de .6 o menores (Aron, 2001).

El uso de la validez implica la existencia de una relación entre la actuación de los sujetos en el test, y otros hechos observables e independientes, pero relacionados con lo que mide el test. Quizás, la mejor definición de validez puede ser la siguiente: "un juicio evaluativo integral del grado, en que la evidencia

empírica y los fundamentos teóricos apoyan la adecuación y conveniencia de las inferencias y acciones basadas en resultados de exámenes y otros modos de evaluación" (Messick, 1989, p. 13). La teoría clásica de los tests establece tres grupos de validez. Validez de contenido, de criterio y de constructo.

En lo que respecta a la validez de contenido, se atiende a través del estudio piloto y el mismo proceso de redacción, traducción inversa y/o análisis formal de los ítems. Para atender a la validez de criterio, se debe verificar si se concretan las relaciones esperadas entre el constructo medido por el test y otras variables. Y por último, pero quizás lo más importante, validez de constructo debe verificarse si: a) se ha hecho un análisis factorial exploratorio y se muestra la matriz de saturaciones/cargas factoriales; b) se especifica el criterio utilizado para determinar el número de dimensiones en el análisis factorial exploratorio; c) se especifican los tamaños de las muestras utilizadas en los análisis factoriales exploratorio; d) se realiza algún análisis factorial confirmatorio para valorar el ajuste del modelo a los datos; e) si se ofrecen datos estadísticos acerca de la validez convergente y discriminante de las puntuaciones obtenidas con el test.

Un último aspecto sobre la métrica de los instrumentos se refiere a las normas y el muestreo a partir del cual se establecen. En primer lugar, es necesario revisar hacia cuál población fue dirigida al construirse. Es decir, cuál ha sido la cultura o grupo al que se ha administrado para crear sus normas. Y si la muestra escogida fue representativa de dicha población.

A partir de los marcos referenciales antes descritos se han creado instrumentos propios o se han realizado procesos de adaptación de instrumentos preexistentes. Entre estos se encuentra la Prueba de Lectura en Grados Tempranos. La misma fue creada para medir lectura en línea base. Los ítems generados corresponden a un modelo latente (lectura simple) con apoyo científico (Hoover & Gough, 1990). Los ítems fueron generados dependiendo de la subprueba y contiene generación aleatoria de pseudopalabras, tomando en cuenta la estructura lingüística del idioma español. Cuenta además con historias cortas originales, con contenido culturalmente relevante, en formatos estandarizados para asegurar que, tanto su longitud como estructura léxica sean apropiadas para el nivel de desarrollo cognitivo del niño.

Para poder dar seguimiento al desempeño lector de estos niños en grados más avanzados, se desarrollan dos pruebas adicionales que permiten medir habilidades de comprensión lectora, más sofisticadas y de acuerdo a la edad del niño o la niña que toma la prueba.

Sistematización y Monitoreo Global

El monitoreo global de un proyecto de cooperación puede ser difícil de llevar, ya que cada indicador suele requerir una base de datos única, así como un levantamiento de datos para las evaluaciones. Esta realidad suele crear bases de datos aisladas y desarticuladas con informaciones diferentes, levantadas y codificadas bajo diferentes rúbricas y estándares. Los procesos de evaluación de calidad de los datos, suelen ser los momentos en los cuales estos desfases llegan a llamar la atención de los implementadores de proyectos. Resultando tardíos para la buena ejecución del proyecto, además de evidenciar ante los evaluadores externos dicha carencia de gestión.

Para evitar el antes mencionado desfase, el Proyecto Leer, como primera acción de su sistema de monitoreo y evaluación, creó un glosario de términos, y así estandarizó la nomenclatura de variables en todas sus bases de datos y documentos oficiales. Se toma el mismo procedimiento para el uso de acrónimos durante la vida del proyecto. Con esta codificación similar, se pueden crear bases conjuntas sin arriesgar errores al momento de transcribir. Las plantillas de digitación se crean de manera estandarizada en Excel, con un formato de plantilla única. La plantilla única permite que, en evaluaciones masivas de lápiz y papel, todo el proceso de digitación se realice con un único formato, única nomenclatura y con bloqueos en celdas que excluyen datos no permitidos por ítem. De igual manera, el proceso de medición y cálculo de cada indicador queda descrito en el Plan de Monitoreo y Evaluación del proyecto, y en su Manual de Procedimientos Operativos Estandarizados.

Luego de dar estos pasos de estandarización, se procede a crear un grupo de bases de datos globales. En ellas, se recoge toda la información relevante de

todas las bases de datos del proyecto, y se agrupan en no más de cuatro bases, permitiendo el cruce de información, así como la fácil estimación de indicadores y generación temprana de informes.

La base de datos global de escuelas, por ejemplo, permite visualizar todas las escuelas del proyecto, su momento de entrada al mismo, cuáles componentes de la intervención recibe, la fecha de inicio de cada componente, el número de estudiantes por año, estudiantes que reciben servicios específicos para discapacidad, y las evaluaciones en las cuales ha participado la escuela. Este manejo permite despersonalizar el proyecto del desempeño particular de un niño o niña, y extenderlo a unidades de intervención más amplias, que nos permiten visualizar cómo características globales de un entorno educativo pueden afectar el desempeño de niños particulares.

Incorporación de la tecnología en el levantamiento de datos

El Proyecto Leer actualmente utiliza una plataforma virtual especializada para levantamiento de datos de lectura. Esta plataforma, desarrollada por RTI International (Pouezevara & Strigel, 2011) se alojó en los servidores de UNIBE. La misma permite utilizar dispositivos electrónicos como medio de registro de información. Los instrumentos de medición creados por el equipo de Monitoreo y Evaluación son adaptados a la plataforma, de manera que se sustituye el uso de papel y lápiz cuando es apropiado. Cada evaluador tiene una tableta Android, con la que registra los datos de múltiples participantes de manera individual, y estos datos se sincronizan a una base de datos única, a través de internet. Esto permite obtener resultados en tiempo real, y reduce, de forma considerable, los errores de procesamiento que son comunes en levantamientos a papel y lápiz. De igual forma, la plataforma facilita la ejecución de entrevistas, ya que contiene un temporizador integrado, y el evaluador no tiene que estar constantemente prestando atención a un material complementario, hoja de respuesta, temporizador, y un participante. De esta manera se reduce la sobre carga cognitiva y los sesgos del evaluador.

Esta plataforma ha sido utilizada también para registrar las observaciones de los docentes, a partir de los momentos de acompañamiento realizado por los mentores del proyecto. En dicho sentido, da seguimiento al proceso de cambio en los docentes de manera continua.

Discusión

La creación de un sistema de monitoreo y evaluación sólida permite la medición de impacto de intervenciones, la toma de decisiones basadas en evidencia y la rápida corrección de aspectos de la intervención que no estén funcionando de la manera esperada. Esta estructura permite, además, flexibilidad y adaptabilidad de las intervenciones a sucesos coyunturales que puedan afectar aspectos del diseño original. La generación de datos confiables en este tipo de sistema, permite además la generación constante de un alto volumen de publicaciones técnicas y científicas que fortalece la literatura científica local.

No podemos subestimar el impacto de este último punto. Solo a través de sistemas rigurosos, con buena instrumentación y procesos con fiabilidad mostrada, podemos apostar a retroalimentar al sistema educativo de forma responsable, asegurando que las recomendaciones realizadas a los tomadores de decisiones, partan de evidencia local de calidad, que se aleja de modelos aspiracionales o de las necesidades reales de los estudiantes dominicanos. Estos sistemas además permiten que la comunidad científica nacional asuma un rol protagónico en la generación de soluciones propias a las problemáticas nacionales.

Preguntas para la discusión en las comunidades de aprendizaje

- ¿Cómo pueden aportar estos sistemas a los proyectos de mi institución?
- ¿Cómo se asegura que no se pierda la información recolectada a través de plataformas tecnológicas en caso de problemas técnicos con los equipos?

Referencias

Aiken, L. R. (2003). *Test psicológicos y evaluación* [11ava. Edición]. México, D.F.: Pearson.

Boyle, G. J., Saklofske, D. H., & Matthews, G. (2014). Criteria for Selection and Evaluation of Scales and Measures. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (pp. 3–15). Elsevier Inc. https://doi.org/10.1016/B978-0-12-386915-9.00001-2

Braden, J. P. (2003). Psychological Assessment in School Settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Volume10 Assessment Psychology* (pp. 261–290). Hoboken, New Jersey: John Wiley & Sons.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216. Recuperado de https://doi.org/10.1177/135910457000100301

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, R. J., & Swerdlik, M. E. (2006). *Pruebas y evaluación psicológicas: Introducción a las pruebas y a la medición (6ta. Ed.)*. México, D.F.: McGraw-Hill Interamericana.

Cronbach, L. J. (1950). Further Evidence on Response Sets and Test Design. *Educational and Psychological Measurement*, *10*(1), 3–31. Recuperado de https://doi.org/10.1177/001316445001000101

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. Recuperado de https://doi.org/10.1007/BF02310555

Díaz Esteve, J. V. (1997). *La Teoría de las Respuestas a los Ítems: Aplicada a la Construcción de los Tests de Aptitudes*. Valencia, España: Cristóbal Serrano Villalba.

Dicken, C. (1963). Good Impression, social desirability, and Acquiescence as suppressor vatriables. *Educational and Psychological Measurement*, *23*, 699–720.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Ft Worth, TX: Dryden Press.

Greenwald, H. J., & Satow, Y. (1970). A Short Social Desirability Scale. *Psychological Reports*, *27*, 131–135. Recuperado de https://doi.org/10.2466/pr0.1970.27.1.131

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing*, *2*(2), 127-160.

Jackson, D. N., & Messick, S. (1962). Response styles on the MMPI: Comparison of Clinical and Normal Samples. *Journal F Abnormal Psychology*, *65*, 285–299.

Jones, P. S., Lee, J. W., Phillips, L. R., Zhang, X. E., & Jaceldo, K. B. (2001). An adaptation of Brislin's translation model for cross-cultural research. *Nursing Research*, *50*(5), 300–304. Recuperado de https://doi.org/10.1097/00006199-200109000-00008

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13–103). New York: Macmillan.

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema*, *28*, 465–470. Recuperado de https://doi.org/10.7334/psicothema2016.113

Nederhof, J. A. (1985). Methods of coping with social desirability: A review. *European Journal of Social Psychology*, *15*, 263–280. Recuperado de https://doi.org/10.1002/ejsp.2420150303

Pouezevara, S., & Strigel, C. (2011). Using Information and Communication Technologies to Support EGRA. *The Early Grade Reading Assessment*, 183.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (Vol. 1, pp. 1–16). Recuperado de https://doi.org/10.1016/B978-0-12-590241-0.50003-4

Sousa, V. D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments

or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice*, *17*(2), 268–274. Recuperado de https://doi.org/10.1111/j.1365-2753.2010.01434.x

Sperber, A. D. (2004). Translation and validation of study instruments for cross-cultural research. *Gastroenterology*, *126*, S124–S128. Recuperado de https://doi.org/10.1053/j.gastro.2003.10.016