Muestreo y análisis de estudios educacionales con R

Andrés Gutiérrez

2017-11-05

Índice general

Capítulo 1

Resumen

Este documento presenta una propuesta de diseño de muestreo que debe aplicarse en el desarrollo del XXXXXXX. En este sentido, el documento hace especial énfasis en la aplicación del piloto en su versión 2018 y en la aplicación del estudio principal para 2019 proporcionando algunos concepto básicos de muestreo, definiendo de manera clara el diseño de muestreo propuesto y mostrando a su vez con ejemplos la utilización de este. Adicional a esto, el documento dedica una sección al tamaño de muestra y se presenta un ejercicio de cómo sería dichos tamaños para los países participantes en el XXXXXXX utilizando estimadores complejos e imputando el tamaño de muestra a los países que no participaron en el estudio. Una de las mayores ventajas de esta propuesta es que el equipo de trabajo cuenta con una vasta experiencia práctica en el desarrollo de metodologías de muestreo y tanto la programación de los algoritmos de muestreo como las referencias bibliográficas de las técnicas utilizadas para la recolección de información primaria son de autoría propia del equipo de trabajo, habiéndose publicado en libros, artículos y paquetes reconocidos y avalados por el CRAN del software estadístico R.

Capítulo 2

Introducción al XXXXXXX 2019

El XXXXXXX (XXXXXXX) es el estudio educativo a gran escala más importante de Latinoamérica y es coordinado por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE[1]), una red de directores nacionales de evaluación educativa de América Latina y el Caribe, coordinada por la Oficina Regional de la UNESCO en Santiago (Chile).

El objetivo de este estudio es brindar información válida y confiable que permita evaluar la calidad de la educación en la región y guiar la toma de decisiones en políticas públicas educativas. Para tal fin, XXXXXXX evalúa el desempeño de los estudiantes en las áreas de lenguaje (lectura y escritura) matemáticas y ciencias, a través de pruebas estandarizadas. Además, aplica cuestionarios de contexto a estudiantes, padres de familia, profesores y rectores, los cuales reúnen información sobre los factores asociados al aprendizaje.

La primera versión del estudio (PXXXXXXX: Primer XXXXXXX) se realizó en 1997 y contó con la participación de 13 países. Este primer estudio evaluó a los estudiantes de tercero y cuarto grado, en las áreas de lectura y matemáticas, y permitió a los países participantes medir el logro de sus estudiantes desde una perspectiva comparativa[2].

En 2006 se aplicó la segunda versión del estudio (SXXXXXXX: Segundo XXXXXXX), la cual no solo reunió más países y economías (17 en total), sino que también introdujo novedades a la prueba: primero, su población objetivo cambió y se enfocó en los estudiantes de tercero y sexto grado; segundo, tuvo un alcance más amplio al incorporar la evaluación de los aprendizajes en escritura y ciencias; terceroo, aumentó el número de contenidos compartidos de los currículos de la región; y cuarto, incluyó preguntas abiertas en matemáticas y ciencias. Todas estas innovaciones hacen que los resultados de los dos primeros estudios no sean comparables directamente, pero hicieron posible un notorio desarrollo de la prueba[3].

Finalmente, la terceroa versión del estudio (TXXXXXXX: tercero XXXXXXX) se llevó a cabo en 2013

[1] ver LLECE
[2] ver PXXXXXXX website
[3] ver SXXXXXXX website

y evaluó las áreas de lenguaje (lectura y escritura) y matemáticas en los grados terceroo y sexto y el área de ciencias naturales en sexto, en 16 países y economías participantes. Este estudio tuvo también algunas novedades: primero, incluyó preguntas sobre el uso de las tecnologías de la información y la comunicación (TIC); segundo, permitió que, en los cuestionarios de contexto, los países y economías participantes agregaran un módulo con preguntas enfocadas en el contexto nacional; terceroo, modificó los criterios de exclusión para que fueran coherentes con los criterios utilizados en otras pruebas internacionales; y cuarto, incorporó la técnica de valores plausibles[4].

Además de estas innovaciones, algunas características de TXXXXXXX que sirven como referente para XXXXXXX 2019 corresponden al diseño de la prueba. Esta aplicación del estudio tuvo una prueba en papel para los estudiantes de terceroo y sexto, e incluyó también cuestionarios para padres de familia, profesores y rectores. Los estudiantes de terceroo contestaron una prueba con 15 preguntas de selección múltiple con una duración de 20 minutos; mientras que los estudiantes de sexto contestaron una prueba con 39 preguntas de selección múltiple con una duración de 45 minutos. Los cuestionarios de padres de familia, profesores y rectores estuvieron compuestos por alrededor de 40 preguntas cada uno[5]. Finalmente, los resultados de las pruebas de lectura, matemáticas y ciencias se presentaron siguiendo una escala con media 700 y desviación estándar 100.

En conjunto, las tres versiones del estudio han sentado las bases para el desarrollo del Cuarto XXXXXXX (XXXXXXX 2019), el cual se enmarca dentro de los mismos lineamientos ya establecidos: evaluará las mismas áreas y grados que TXXXXXXX y hará uso de una prueba estandarizada en papel y la aplicación de los distintos cuestionarios de contexto mencionados.

Este documento presenta entonces una propuesta muestreo para la aplicación del estudio piloto de XXXXXXX 2019, donde el objetivo principal será evaluar la validez y pertinencia de los ítems. Además, esta es una propuesta para participar activamente en el desarrollo de todas las etapas del estudio y garantizar su ejecución con altos estándares de calidad.

Una ventaja técnica de esta propuesta es que se usaron las bases de datos[6] del TXXXXXXX, disponibles en el sitio web de la Unesco, para realizar los cálculos pertinentes a configurar una propuesta más cercana a la realidad del estudio. Todos estos cómputos fueron realizados en el software estadístico `R`, con librerías desarrolladas por el equipo de expertos que presentan esta propuesta, y avaladas directamente por el `Comprehensive R Archive Network - CRAN`.

[4] ver TXXXXXXX website

[5] ver: UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX pg. 197

[6] ver Bases de datos

Capítulo 3

Definiciones básicas del muestreo

El muestreo en la investigación educativa se realiza generalmente con el fin de permitir el estudio detallado de un subconjunto representativo de la población. La información derivada de la muestra resultante se emplea habitualmente para desarrollar generalizaciones útiles sobre la población. Estas generalizaciones pueden ser estimaciones de una o más características asociadas con la población o pueden estar relacionadas con estimaciones de la fuerza de las relaciones entre las características dentro de la población.

Siempre que se utilicen procedimientos científicos de muestreo, la selección de una muestra a menudo ofrece muchas ventajas en comparación con una cobertura completa de la población. Por ejemplo, la reducción de los costos asociados con la recolección y el análisis de los datos, la reducción de los requisitos para el personal capacitado para llevar a cabo el trabajo de campo, la velocidad mejorada con mayor precisión debido a la posibilidad de una supervisión más intensa del trabajo de campo y las operaciones de preparación de datos.

3.1 Descripción de la población de interés

Para XXXXXXX las poblaciones de interés son los estudiantes que se encuentren en los grados terceroo y sexto. La base de comparación entre los países participantes será el número de años de educación formal que los estudiantes hayan recibido. De acuerdo a la Clasificación Internacional Normalizada de la Educación (CINE), el nivel CINE 1 corresponde a la educación primaria y el primer año escolar en este nivel marca el comienzo de la educación básica. Así, los grados objetivo de XXXXXXX se encuentran tres y seis años después de este año escolar, los cuales corresponden a los grados terceroo y sexto en la mayoría de países.

De esta manera, se espera que todos los estudiantes en los grados terceroo y sexto tengan tres y seis años de escolaridad, respectivamente; además, la edad promedio de los estudiantes debería ser 8,5 años en terceroo y 11,5 años en sexto. Sin embargo, existen diferencias entre los países en cuanto a la definición de los grados escolares y la edad promedio de los estudiantes. Por esta razón, es muy

importante que los países participantes brinden información clara y completa sobre estas variables al momento de comenzar con las tareas de muestreo.

Una vez definidos los grados de interés, es importante considerar las características que debe tener una escuela para ser parte del estudio:

- Ser reconocida por una autoridad competente en el país e impartir educación en los grados de interés.
- Tener una ubicación física y un sistema de administración únicos.
- Tener un número definido e identificable de estudiantes y profesores.
- Tener un espacio social identificable.

Cabe resaltar que en algunas ocasiones las escuelas no cumplen todos los criterios, por lo que es necesario analizar todos los casos que se presenten. Por ejemplo, si dos escuelas comparten una misma ubicación física, pero tienen horarios, estudiantes y personal educativo distinto, se consideran como escuelas diferentes. De igual forma, dos sedes de una misma escuela que tengan ubicaciones físicas distintas y que tengan, por lo tanto, estudiantes y personal educativo distinto, son consideradas como escuelas diferentes. Finalmente, en el caso de las escuelas multigrado, se considerará únicamente a los estudiantes que, dentro del aula multigrado, se encuentren en grados terceroo o sexto[1].

Dentro de las escuelas que cumplan estos criterios serán seleccionados los estudiantes que harán parte del estudio. Estos estudiantes deben estar matriculados en el sistema educativo formal (en los grados de interés, sin importar su edad) y considerarse aptos para responder las pruebas. Para ser considerado apto, un estudiante debe estar en la capacidad física y cognitiva de contestar una prueba de logro de aprendizaje en el idioma oficial del país participante.

En línea con lo anterior, no harán parte de la población de interés de XXXXXXX: escuelas donde se imparta educación no formal, estudiantes fuera del sistema educativo formal (aquellos que toman clases de tiempo completo en su hogar, por ejemplo), estudiantes con necesidades especiales y escuelas especializadas en este tipo de estudiantes, escuelas para adultos, y escuelas donde se enseñe solo un idioma y este sea diferente al oficial del país participante.

Será responsabilidad de cada país brindar información precisa sobre las escuelas y estudiantes elegibles para XXXXXXX. Cada país participante deberá identificar las escuelas que tengan estudiantes en los grados de interés y que cumplan con los demás criterios mencionados, las cuales deben incluirse en un listado sin datos incorrectos o duplicados. Además, es importante que el país informe sobre los posibles inconvenientes que se puedan presentar con la población objetivo.

Las fechas de aplicación de la prueba tienen que ser apropiadas para la población objetivo: no pueden estar dentro de las primeras seis semanas del año escolar, debido a la preocupación de que el desempeño de los estudiantes pueda ser más bajo en este periodo (incluso después de controlar por la edad); no debe cubrir un periodo mayor a 42 días (a menos que eso sea acordado con el país participante); y deben considerarse los calendarios escolares (norte y sur).

[1] ver UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX pg. 200

3.2 Descripción de las subpoblaciones de interés

Las subpoblaciones de interés o estratos son divisiones de la población objetivo en grupos que comparten una característica común y que son importantes para el país. Por esta razón, estas agrupaciones se definen en conjunto con el país participante y, al hacerlo, se tienen en cuenta variables de estratificación explícita e implícita.

La estratificación explícita consiste en agrupar las escuelas en estratos que serán tratados de forma independiente (como si se tuvieran marcos de muestreo separados para cada subpoblación), de acuerdo a una variable específica. Mientras que la estratificación implícita consiste en clasificar las escuelas dentro de cada estrato explícito haciendo uso de un conjunto de variables definidas[2]. A continuación se ofrecen algunos ejemplos para cada una de estas técnicas.

3.2.1 Estratificación explícita

Los estratos explícitos son útiles para reducir la varianza de muestreo y asegurar la representatividad de los estudiantes en cada uno de los grupos de escuelas que comparten una característica común[3]. Algunas variables que se consideran en el proceso de estratificación explícita son:

- Estados o regiones de un país.
- Zona en la que está ubicada la escuela: urbana o rural. Aquí cada país brinda su definición de ruralidad.
- Dependencia administrativa: gestión pública o privada.
- Grado escolar: solo terceroo, solo sexto, ambos grados.

Los países participantes pueden proponer otras variables de interés para considerar como variables de estratificación explícita, las cuales serán analizadas cuidadosamente entre cada Coordinador Nacional, el Comité de Datos y la Coordinación en UNESCO-OREALC.

3.2.2 Estratificación implícita

Este tipo de estratificación es una forma de garantizar una asignación estrictamente proporcional de las escuelas en todos los estratos implícitos. También puede conducir a una mayor confiabilidad de las estimaciones del estudio, siempre que las variables de estratificación implícita que se consideran estén correlacionadas con el logro de aprendizaje evaluado en el nivel de la escuela . Algunas variables de estratificación implícita son:

- Género: si un país tiene escuelas solo para hombres o solo para mujeres, es importante incluir esta variable para evitar desequilibrios por género.
- Composición de las minorías: escuelas con grupos minoritarios (grupos étnicos, por ejemplo).

[2] ver PISA (2012), Technical Report. pg. 71
[3] ver PISA (2012), Technical Report. pg. 81

- Jornada de la escuela: en caso de que la escuela tenga jornada diurna, nocturna, completa, etc.
- Tamaño de la escuela: medida con el número de estudiantes en los grados de interés, es la última variable de estratificación que se aplica en el marco de muestreo (medida de tamaño).

Al igual que con las variables de estratificación explícita, los países participantes pueden proponer otras variables de interés para considerar como variables de estratificación implícita, las cuales serán analizadas cuidadosamente. Por ejemplo, un país puede estar interesado en particionar las escuelas pertenecientes a su sistema educativo con respecto al número de estudiantes adscritos (matrícula) en las municipalidades (o divisiones administrativas), de la siguiente manera:

- *Municipios capitales*: que son las ciudades capitales de departamentos o estados (mayor desagregación geográfica).
- *Municipios grandes*: referentes a municipios con matrícula mayor a 10.000 estudiantes.
- *Municipios intermedios*: que son municipios con matrícula entre 5.001 y 10.000 estudiantes.
- *Municipios medianos*: que representan a los municipios con matrícula entre 2.501 y 5.000 estudiantes.
- *Municipios pequeños*: con matrícula menor a 2.500 estudiantes.

3.2.3 Dominios poblacionales

Definidos por aquellos subgrupos de interés que se conocen una vez se haya aplicado el instrumento. Los dominios poblacionales son subgrupos de interés que cumplen las siguientes características:

- Ningún elemento de la población pertenece a dos dominios.
- Todo elemento de la población pertenece a un único dominio.
- La población del estudio puede representarse como la unión de todos los dominios[4].

Por ejemplo, entre los dominios poblacionales podemos encontrar subgrupos de la población divididos según género, calendario escolar, entidades territoriales de cada país, jornadas escolares o programas de mejora de la calidad educativa focalizados.

Cabe resaltar que la identificación de las unidades de la población (escuelas y estudiantes) y su posterior clasificación en un dominio, solo se puede realizar cuando los países participantes han entregado toda la información requerida y el proceso de medición ha comenzado.

3.2.4 Post-estratos

Para calibrar los pesos de muestreo y que las estimaciones sean más precisas, es decir, que tengan un menor margen de error, se hace uso de post-estratos para escuelas y estudiantes. Los post-estratos son subgrupos poblacionales de los que se conoce su tamaño, pero no se sabe cuál será su número real de individuos en la muestra realizada. Esto sucede en dos casos:

[4] ver Gutiérrez, A. (2015), pg. 83

1. Todas las unidades poblacionales (escuelas y estudiantes) están clasificadas en los subgrupos según el marco de muestreo; sin embargo, por facilidad logística se decide no usar esta información. Esta característica se vuelve a observar solo después de haber seleccionado la muestra.
2. Los tamaños de los subgrupos poblacionales se conocen, pero no es posible clasificar las unidades poblacionales en ellos debido a falencias en el marco de muestreo. Esto se soluciona después de haber seleccionado la muestra.

En los dos casos anteriores, la información sobre la clasificación de escuelas y estudiantes en los post-estratos que se obtiene después de tener la muestra, se utiliza para mejorar la eficiencia de los estimadores. En general, el uso de los post-estratos es útil para calibrar los pesos de muestreo de las escuelas (por número de establecimientos, establecimientos por zona/sector) y para calibrar los pesos de muestreo de los estudiantes (número total de estudiantes a nivel nacional, estudiantes por género o estudiantes por zona/sector).

3.2.5 Sobremuestras de interés

Una sobremuestra es una muestra adicional de la población objetivo para un subgrupo particular como una ciudad o una región. Cada país participante puede decidir si tener una o varias sobremuestras, dependiendo del interés que tenga en una subpoblación específica.

Sin embargo, es importante resaltar que esta es una negociación que deben tener los coordinadores nacionales con el socio implementador, puesto que, si bien contar con una sobremuestra brinda mayor información sobre un subgrupo de interés, esta opción incrementa el tamaño de muestra y los costos asociados al estudio. Cada país puede manifestar su interés en tener una o más sobremuestras y el socio implementador entregará diferentes escenarios de muestreo de acuerdo a estos requerimientos.

TXXXXXXX[5] consideró tres casos en los que tener una sobremuestra puede ser recomendable para un país:

1. Si estaba participando en los Módulos Nacionales del TXXXXXXX, puesto que los módulos podían contener factores asociados al aprendizaje específicos para el país, especialmente relacionados con algún subgrupo poblacional.
2. Cuando el error muestral de los estratos o subpoblaciones de interés supera un valor significativo (más de 10 puntos de la prueba), es decir, cuando el tamaño de algún estrato explícito es muy pequeño y no permite realizar análisis estadísticos precisos.
3. Cuando existe un interés por parte del país para realizar análisis precisos sobre un subgrupo poblacional particular.

En evaluaciones internacionales como PISA la selección de sobremuestras se realiza porque el país está interesado en algún subgrupo poblacional. En este caso, se agrega una nueva variable de estratificación explícita y se generan los estratos explícitos correspondientes. En general, los países pueden considerar tener una sobremuestra si están participando en los *Módulos Nacionales del XXXXXXX*, el

[5] ver UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX, págs. 211 y 212

error muestral de los estratos o subpoblaciones de interés supera un porcentaje significativo, o desea realizar un análisis sobre un subgrupo de la población.

3.2.6 Medidas de tamaño

Definir una medida adecuada de tamaño (MOS) es un aspecto crítico en el desarrollo del plan de muestreo para cada país, puesto que el tamaño de la escuela determina su probabilidad de selección. La medida de tamaño escolar más adecuada es un recuento actualizado del número de estudiantes en los grados objetivo. Si el número de estudiantes en estos grados no está disponible, la matrícula total de la escuela ofrece el mejor valor aproximado[6].

El número de estudiantes en los grados objetivo es la medida de tamaño por excelencia utilizada en distintas evaluaciones internacionales. Por ejemplo, PISA solicita el conteo de estudiantes de 15 años en cada escuela, TIMSS el conteo de estudiantes en los grados cuarto y octavo, PIRLS el conteo de estudiantes en cuarto grado, e ICCS el conteo de estudiantes de 14 años.

Para XXXXXXX los grados de interés son terceroo y sexto, por lo que la medida de tamaño será la matrícula en estos grados para cada escuela. En TXXXXXXX[7] se utilizaron tres variables para calcular la medida de tamaño: la matrícula en terceroo, la matrícula en sexto, y la suma de ambas matrículas, según el estrato al que pertenecía la escuela.

3.3 Marco de muestreo

El marco de muestreo es una lista de todas las escuelas del país que cumplen las condiciones para ser parte del estudio y que tienen estudiantes matriculados en los grados objetivo. Esta es la lista de la cual se seleccionará la muestra de escuelas que participarán en el estudio. Se espera que este marco incluya cualquier escuela que pueda tener estudiantes matriculados en terceroo y sexto, incluso aquellas escuelas que podrían ser excluidas o consideradas inelegibles (porque en el momento de la aplicación no tienen estudiantes que cumplan todos los requisitos, por ejemplo).

El marco muestral es, por lo general, una hoja de cálculo que contiene una sola entrada para cada escuela. Esta entrada incluye un número de identificación único, los valores de las variables de estratificación para la escuela y la medida de tamaño escolar previamente definida. Si no existieran estimaciones razonables para la medida de tamaño escolar o si los datos disponibles estuvieran desactualizados, las escuelas tendrían que ser seleccionadas con probabilidades iguales, lo que podría requerir un aumento en el tamaño de la muestra[8]. Cabe mencionar que si el plan de muestreo requiere una estratificación explícita, debería existir un marco de muestreo separado para cada estrato explícito, o equivalentemente una variable que indique el estrato al que pertenece cada escuela.

[6] ver Joncas, M. & Foy, P. (2013), pg. 10

[7] ver UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX, pg. 214

[8] ver Joncas, M. & Foy, P. (2013), págs. 10 y 11

Una vez los países envían el marco de muestreo, es necesario realizar una serie de validaciones. Por ejemplo, verificar que los datos estén actualizados, que la información coincida con cualquier reporte que el país haya realizado previamente y con los datos de las aplicaciones anteriores del estudio, revisión de escuelas excluidas, de valores duplicados y datos faltantes.

Esta tarea es muy importante, puesto que un marco de muestreo bien construido y que no esté contaminado con datos incorrectos o duplicados, proporciona una cobertura completa de la población objetivo, lo cual afecta directamente las probabilidades de selección de las escuelas, sus pesos, las estimaciones finales y, en general, los resultados del estudio.

Por último, una vez que las escuelas sean seleccionadas y accedan a participar en el estudio, deben enviar un listado actualizado con todos los estudiantes elegibles (todos los estudiantes en grados tercero y sexto). Este listado debe tener un número de identificación para cada estudiante y especificar para cada uno el género, la edad, el grado al que pertenece y si tiene algún tipo de discapacidad que le impida presentar la prueba.

3.3.1 Unidades de muestreo

En esta propuesta se consider que la unidad primaria de muestreo (UPM) son las escuelas (en la primera etapa del muestreo). Las unidades secundarias de muestreo (USM) son los estudiantes (en la segunda etapa del muestreo). Estas serán las unidades de selección establecidas para todos los países, para efectos de comparabilidad. Sin embargo, en algunos casos esta unidad puede cambiar teniendo en cuenta la estructura particular del sistema educativo, al estructura de las escuelas y de los programas escolares.

3.3.2 Unidades de observación

La unidad de observación para XXXXXXX son los estudiantes, aquellos que son seleccionados para presentar la prueba. De estas unidades de observación se desprenderá el análisis de ítems que permitirá establecer la viabilidad de los mismos en el estudio principal. Nótese que no sólo se trata de validar los instrumentos cognitivos. Por ejemplo, el TXXXXXXX contó con cuestionarios para estudiantes, familias, profesores y directores. La información consultada mediante estos instrumentos hizo posible realizar análisis de factores asociados respecto de las características principales de los sistemas educativos participantes.

3.3.3 Unidades de análisis

Corresponden a los diferentes niveles de desagregación establecidos por los países para consolidar el diseño probabilístico y sobre los que se presentan los resultados de interés. Adicional a esto, es necesario tener en cuenta que la unidad principal de análisis es el país.

3.3.4 Periodo de referencia

Según la necesidad establecida en el pliego de condiciones donde se afirma que:

> *"De acuerdo con el cronograma técnico del XXXXXXX-2019, durante los años 2017 y 2018 se encuentran previstos dos grandes procesos para la consecución del estudio. Durante el primer año, se considera la preparación de los marcos de referencia y el diseño de los instrumentos de evaluación, para las pruebas de logros de aprendizaje y de factores asociados. El proceso de pilotaje de instrumentos se realizará en el año 2018, todo como preparación para la aplicación definitiva prevista para el año 2019, y la respectiva entrega de resultados en 2020."*

Luego, el periodo de referencia del estudio piloto está dado para el año 2018.

3.4 Exclusiones

Al aplicar un estudio como XXXXXXX es necesario considerar algunos escenarios en los que un grupo de escuelas o estudiantes es excluido. Por ejemplo, puede ser que un país requiera excluir una región geográfica pequeña por los problemas logísticos que una aplicación en esa zona pueda generar (zonas de difícil acceso), o que deba excluir un grupo étnico por razones políticas o porque sus miembros no hablan el idioma oficial en el que se aplicará la prueba.

Todos estos casos en los que es posible realizar algún tipo de exclusión deben ser definidos para todos los países participantes, teniendo en cuenta que las tasas de exclusión deben ser limitadas. En otras palabras, se debe evitar excluir una proporción significativa de escuelas o estudiantes, puesto que afectaría los resultados del estudio y no serían representativos de todo el sistema escolar del país. En conjunto, se espera que las exclusiones a nivel de escuelas y estudiantes no superen el 5% de la población objetivo.

Todas las exclusiones presentadas por el país deben ser analizadas para garantizar que no superan las proporciones establecidas, que no afectan la representatividad, que se trata de cifras actuales y reales (se sugiere que el país comparta, en la medida de lo posible, sus fuentes de información), que estuvieran dentro de las exclusiones definidas, y que sean coherentes con los datos de las aplicaciones anteriores del estudio.

Las exclusiones adicionales propuestas por el país pueden ser negociadas, siempre y cuando no superen el porcentaje de exclusiones permitido. Por ejemplo, en PISA es posible considerar excluir las escuelas con menos de tres estudiantes elegibles, si el país no ha superado la tasa de exclusión escolar del 2,5%; además, si estas escuelas representan menos del 0,5% de estudiantes, si estas exclusiones se han realizado en aplicaciones anteriores y si no ocasionan la pérdida de estratos enteros, se pueden excluir[9].

[9] ver PISA (2012), Technical Report, pg. 82

3.4.1 Exclusión a nivel de escuela

La propuesta para las categorías de exclusión a nivel de escuela a ser consideradas en XXXXXXX son:

- Escuelas de difícil acceso debido a su ubicación geográfica remota. Estas no deben superar el 0,5% del número total de estudiantes.
- Escuelas pequeñas con menos de cuatro estudiantes en los grados de interés. Estas no deben superar el 2% del número total de estudiantes. Cabe resaltar que en ese 2% se deben excluir las escuelas más pequeñas entre las que tengan menos de cuatro estudiantes.
- Escuelas que imparten educación únicamente a estudiantes con necesidades educativas especiales, con discapacidades intelectuales o funcionales, o con experiencia limitada en el lenguaje de evaluación. Estas no deben superar el 2% del número total de estudiantes.
- Escuelas con una estructura o currículo radicalmente diferente al sistema educativo convencional, identificadas a priori por los Coordinadores Nacionales del XXXXXXX en cada país.

Con respecto a las escuelas pequeñas, es importante mencionar que esta es una característica que puede relacionarse con otras variables como ruralidad o escuelas multigrado. Al excluirlas, se debe tener cuidado de no reducir considerablemente el tamaño de estas subpoblaciones de interés. De igual forma, en países caracterizados por escuelas pequeñas, excluirlas causaría una pérdida de representatividad[10].

3.4.2 Exclusión a nivel de estudiante

La propuesta de categorías de exclusión a nivel de estudiante a ser consideradas en XXXXXXX son estudiantes con discapacidades o con insuficiencia en el idioma de la prueba. Estas son las categorías utilizadas en evaluaciones internacionales como PISA, TIMSS y PIRLS .

- *Estudiantes con discapacidades funcionales*: estudiantes con discapacidades físicas permanentes que les impiden presentar la prueba. Aquellos estudiantes con algún tipo de discapacidad física que no represente un impedimento para ser evaluados deben presentarla.
- *Estudiantes con discapacidad intelectual*: estudiantes que han sido evaluados por personal médico y cuyo diagnóstico haya sido discapacidad intelectual, o que, según la opinión profesional del director de la escuela o de otros miembros calificados de la escuela tienen un retraso cognitivo tal que no pueden ser evaluados válidamente. Esto incluye a los estudiantes que son emocional o mentalmente incapaces de seguir incluso las instrucciones generales de la prueba. Cabe resaltar que los estudiantes no deben ser excluidos únicamente por mal desempeño académico o por problemas disciplinarios.
- *Estudiantes con experiencia insuficiente en el idioma de la prueba*: estudiantes con una experiencia lingüística insuficiente que no pueden leer o hablar el idioma de la prueba. Estos estudiantes no son hablantes nativos del idioma de evaluación, tienen una competencia limitada en el mismo y han recibido menos de un año de instrucción en dicho lenguaje.

[10] ver UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX, pg. 204

Debido a que los criterios de discapacidad varían de un país a otro, es importante que los países consideren las normas internacionales que rigen la inclusión escolar de las personas en condición de discapacidad mental o física. En caso de que sea necesario excluir un grupo de estudiantes que no se haya definido previamente (por ejemplo, estudiantes con otro tipo de trastornos de aprendizaje como dislexia), debe analizarse la posibilidad de agregar una categoría de exclusión adicional.

Todas estas exclusiones deben ser debidamente documentadas, indicando el número de estudiantes en la población objetivo y el número de estudiantes excluidos a nivel de escuela.

3.5 Tasas de participación y cobertura

Las tasas de participación aceptables deben definirse tanto para escuelas como estudiantes, a continuación presentamos una revisión de las tasas definidas para algunos estudios internacionales.

Tasa de respuesta para las escuelas

Por ejemplo, en PISA la tasa de respuesta requerida para las escuelas originalmente muestreadas es del 85%, la cual podía estar entre 65% y 85% al considerar los reemplazos. Aquí es importante tener en cuenta que una escuela es considerada participante si por lo menos el 50% de los estudiantes que debían presentar la prueba, efectivamente la presentan. Mientras que en TXXXXXXX, la tasa de respuesta requerida fue del 80%, donde por lo menos el 70% de las escuelas debieron ser las originalmente muestreadas.

Tasa de respuesta para los estudiantes

En PISA, la tasa de respuesta ponderada global requerida para los estudiantes es 80%. Si la tasa de participación de los estudiantes dentro de una escuela está entre el 25 y el 50%, entonces esa escuela no es considerada para calcular las tasas de respuesta, pero si hace parte de la base de datos completa del estudio. Si la tasa de participación de estudiantes no alcanza el 25% entonces esa escuela no es considerada ni para los cálculos, ni en la base de datos. Por su parte, TXXXXXXX requirió de una tasa de participación de estudiantes igual o mayor al 80%.

Tasa de respuesta conjunta

Evaluaciones internacionales como TIMSS e ICCS proponen tasas de participación en las que el criterio de validez se debe cumplir para escuelas y estudiantes para que sean aceptables. Por ejemplo, en TIMSS para que la muestra sea aceptable debe cumplir una de las siguientes condiciones:

1. Tener una tasa de participación escolar mínima del 85% para escuelas originalmente muestreadas; una tasa mínima de participación en el aula del 95% para las escuelas originalmente muestreadas y las escuelas de reemplazo; y una tasa de participación mínima de estudiantes del 85% para las escuelas originalmente muestreadas y las escuelas de reemplazo.

2. Tener un mínimo combinado de escuela, aula y tasa de participación de estudiantes del 75%, para escuelas originalmente muestreadas (aunque las tasas de participación en el salón de clases pueden incluir escuelas de reemplazo).

Por otro lado, en ICCS se han establecido las siguientes tres categorías, teniendo en cuenta que una escuela es considerada participante si al menos el 50% de sus estudiantes contesta la prueba.

1. Categoría 1: Tasa de participación satisfactoria sin el uso de escuelas de reemplazo. Es aceptable si se cumple uno de los siguientes criterios:

- 85% de participación en la escuela y 85% de la tasa de respuesta de los estudiantes (los porcentajes se redondean al número entero más cercano, ponderado o no ponderado)
- 75% combinando escuela y estudiantes

2. Categoría 2: Tasa de participación muestral satisfactoria con el uso de escuelas de reemplazo:

- Sigue los mismos criterios de la categoría 1, pero incluidas las escuelas de reemplazo y dado que al menos el 50% de las escuelas originalmente muestreadas participaron.
- El reporte incluye anotaciones de sesgo potencial

3. Categoría 3: No se cumplieron los requisitos anteriores, pero cumplen con los procedimientos de muestreo

- Los resultados aparecerán en una sección separada de las tablas para publicación.

Propuesta para el XXXXXXX

La propuesta para XXXXXXX incluye cuatro categorías para las dos clasificaciones establecidas: escuela y estudiantes.

1. Tasa de respuesta satisfactoria: más del 85% de respuesta antes de reemplazos.
2. Tasa de respuesta aceptable: entre el 70% y 85% de respuesta antes de reemplazos y más del 85% de respuesta después de reemplazos.
3. Tasa de respuesta intermedia: entre el 70% y 85% de respuesta antes de reemplazos y entre el 70% y 85% de respuesta después de reemplazos.
4. Tasa de respuesta no aceptable: Menos del 70% de respuesta antes de reemplazos y menos del 70% de respuesta después de reemplazos.

La siguiente figura muestra la distribución sugerida para la categorización del levantamiento de la información, según las tasas de participación en cada país.

3.6 Parámetros de interés

Al aplicar pruebas estandarizadas para medir el logro de aprendizaje de los estudiantes se obtiene un conjunto de respuestas a los ítems evaluados. Con esta información, los parámetros de interés en el estudio piloto del XXXXXXX deben relacionarse con la dificultad de los ítems y su adecuación al constructo que se intenta medir. Con base en lo anterior, también se busca estimar la habilidad de los estudiantes de cada país participante, puesto que se trata de características no observadas. Este proceso de estimación puede realizarse a través de distintos métodos, entre ellos el uso de valores

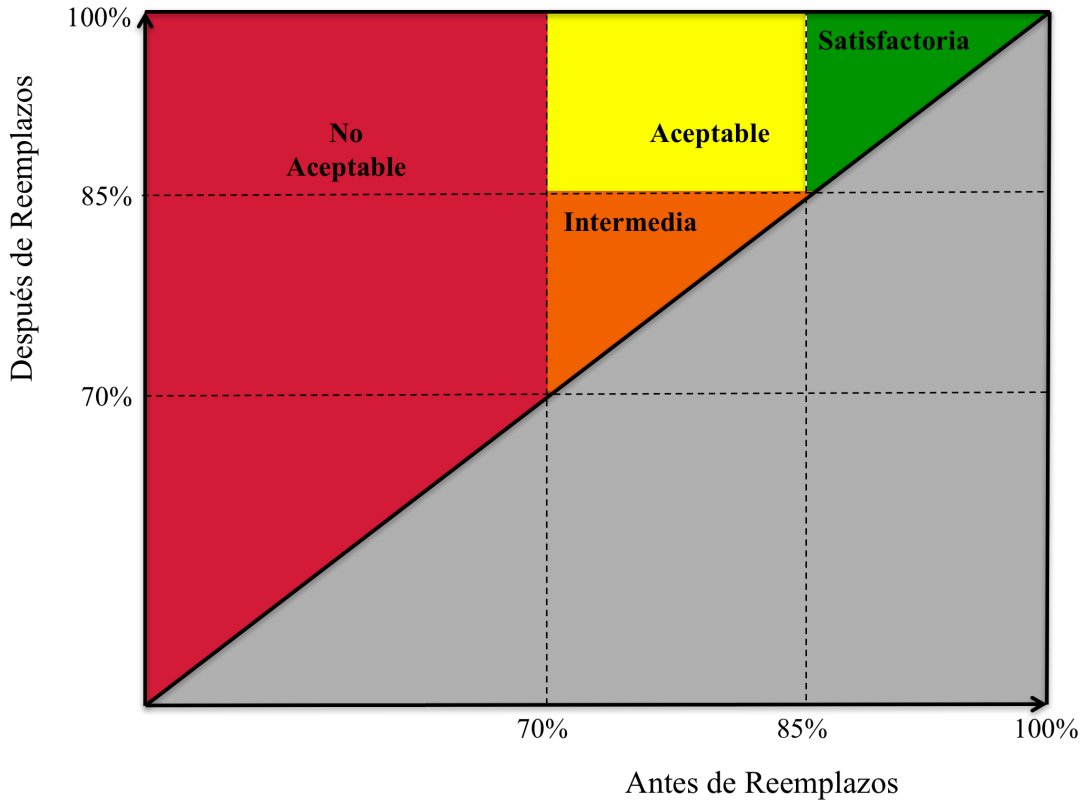

Figura 3.1: Categorización propuesta según las tasas de participación para XXXXXXX

plausibles, definidos como una imputación aleatoria para cada estudiante, la cual indica el nivel probable de desarrollo de las competencias evaluadas. Estos valores surgen de la distribución de las puntuaciones de los estudiantes y contienen un error aleatorio. Los valores plausibles son adecuados para describir el desempeño de la población en su conjunto, puesto que producen estimadores consistentes de los parámetros poblacionales[11].

En general, este documento considera los siguientes parámetros poblacionales de interés como estructura de la estrategia de estimación:

- **Total poblacional**: se define como la suma de las observaciones de la variable de interés en la población. Se calcula mediante la siguiente ecuación:

$$t_y = \sum_U y_k$$

En donde U hace referencia al universo de estudio e y_k hacer referencia a la variable de interés en el k-ésimo individuo.

- **Promedio poblacional**: se define como la suma de las observaciones de la variable de interés en la población dividida por el tamaño poblacional N. Se calcula mediante la siguiente ecuación:

$$\bar{y}_U = \frac{t_y}{N}$$

- **Proporción poblacional**: es un promedio sobre una variable dicotómica z_k que toma el valor de 1 si el k-ésimo individuo tiene el atributo de interés y de 0 en otro caso. Se calcula mediante la siguiente ecuación:

$$P_U = \frac{\sum_U z_k}{N}$$

- **Razón poblacional**: se calcula como el cociente de dos totales, el primer total asociado a una variable de interés y, el segundo total asociado a una variable de interés $\mathbf{Z}$. Se calcula mediante la siguiente ecuación:

$$R_U = \frac{t_y}{t_z}$$

en donde t_y es el total poblacional asociado a la variable y, t_z es el total asociado a la variable $\mathbf{Z}$.

[11] ver PISA (2012), Technical Report, págs. 146 – 147

3.6.1 Parámetros con información de estudiantes

La información de los estudiantes permite identificar las diferencias que existen entre los países en cuanto a la relación que tienen algunos factores de contexto (a nivel de estudiantes, escuelas, etc.) con el logro académico. De igual forma, con esta información es posible estimar la medida en que las escuelas refuerzan o disminuyen los efectos de los factores que, a nivel de estudiantes, están asociados con el desempeño. Finalmente, es posible hacer seguimiento a los cambios que estas variables tienen a través del tiempo, realizando comparaciones con las versiones anteriores del estudio.

Algunos ejemplos de estos parámetros son:

1. Desempeño medio nacional en matemáticas, ciencias, lectura y escritura (puntaje promedio).
2. Desempeño medio por subpoblación de interés en matemáticas, ciencias, lectura y escritura (puntaje promedio).
3. Porcentaje de estudiantes por país categorizados en un nivel de desempeño particular.
4. Porcentaje de estudiantes por subpoblación de interés categorizados en un nivel de desempeño particular.
5. Diferencias en los resultados entre los países .
6. Porcentaje de estudiantes en una categoría del nivel socioeconómico del XXXXXXX.

Es importante mencionar que los resultados más importantes acerca de los logros de aprendizaje de los estudiantes son el **puntaje promedio** y los **niveles de desempeño**. El puntaje promedio es una medida cuantitativa del desempeño de los estudiantes y tiene una escala con una media centrada en 700 y desviación estándar 100. Los niveles de desempeño, por otro lado, son una medida cualitativa que indica lo que los estudiantes saben o son capaces de hacer de acuerdo al nivel en que son ubicados.

3.6.2 Parámetros con información de docentes y escuelas

La información de docentes y escuelas permite identificar las diferencias que existen entre los países en cuanto a la relación que tienen los factores a nivel de escuela y aula con el logro académico. Asimismo, es posible analizar el porcentaje de la varianza que es explicada por diferencias entre o dentro de las escuelas y cómo esa proporción cambia entre países. Finalmente, al igual que con la información por estudiante, es posible hacer seguimiento a los cambios que estas variables tienen a través del tiempo.

Capítulo 4

Diseño de muestreo

El diseño muestral del XXXXXXX debe mantener una estructura similar a la del TXXXXXXX (para conservar los criterios de comparabilidad entre las poblaciones objetivo de ambos estudios) y garantizar que la muestra seleccionada para cada país sea representativa de las dos poblaciones de interés: estudiantes en grados terceroo y sexto.

De esta manera, el diseño de muestreo que se presenta en esta sección sigue dos etapas: en la primera se seleccionan las escuelas que participarán en el estudio y en la segunda se escogen los estudiantes de terceroo y sexto que presentarán la prueba.

4.1 Diseño de muestreo general

XXXXXXX mantendrá un diseño de muestreo similar al de TXXXXXXX, estudio en que se siguió un método de selección estratificado, por conglomerados y bietápico: en la primera etapa se seleccionan las escuelas (dentro de cada estrato explícito, siguiendo un muestreo sistemático y teniendo en cuenta su medida de tamaño) y en la segunda etapa se escoge aleatoriamente un aula dentro de las escuelas previamente seleccionadas.

Cabe mencionar que en el XXXXXXX se selecciona una sola muestra de escuelas y no dos muestras independientes para cada uno de los grados de interés. Esto puede presentar algunas dificultades, puesto que la selección de estudiantes puede resultar desbalanceada, es decir, puede ser que la muestra de estudiantes en un grado sea muy grande y en el otro grado muy pequeña. Para que esto no suceda, en el TXXXXXXX se definieron tres estratos según la presencia de los grados de interés en la escuela: escuelas solo con terceroo, escuelas solo con sexto y escuelas con ambos grados.

Finalmente, es importante tener en cuenta que cada país participante debe tener un plan de desarrollo de actividades que le permita seguir los lineamientos del diseño de muestreo desde la identificación de la población objetivo hasta la selección de una muestra representativa. Este plan de muestreo debe acordarse con el consorcio de investigación, para garantizar que se ajusta a todas las normas

de muestreo.

Además, cada país es responsable de entregar información completa y confiable sobre: la población objetivo, las escuelas que tienen los grados de interés (marco de muestreo), las variables de estratificación explícita e implícita, las exclusiones, entre otras; todos estos datos deben ser debidamente diligenciados en los formatos establecidos. Asimismo, luego de la selección de la muestra los países deben contactar a las escuelas que han sido escogidas, tratar de asegurar su participación y hacer seguimiento a todo el proceso de aplicación.

Diseño de muestreo propuesto

El tipo de muestreo que se considerará en esta propuesta es probabilístico, estratificado y bietápico. El muestreo es probabilístico debido a que las unidades de muestreo en cada una de las etapas tienen una probabilidad conocida y mayor que cero de ser seleccionadas. Es estratificado porque se consideran particiones poblacionales, definidas como conjuntos de elementos cuya unión conforma el universo, que no se traslapan y donde todos son diferentes de vacío, esto tratando de lograr homogeneidad dentro de ellas y heterogeneidad entre ellas. Por último, es bietápico puesto que se considera primero la selección de escuelas en una primera etapa y luego la selección de alumnos en una segunda etapa.

Este diseño de muestreo estima el total de cada escuela t_i mediante una sub-muestra seleccionada desde el marco de muestreo de UPM (escuelas). Suponga que la población de estudiantes U se divide en N_I escuelas, que definen una partición de la población, llamados también **conglomerados** y denotadas como $U_I = \{U_1, \ldots, U_{N_I}\}$ (U_I es la población de todas las escuelas en un país y N_I es el número total de escuelas dentro del país). Note que la i-ésima escuela U_i $i = 1, \ldots, N_I$ contiene N_i estudiantes. Luego, el proceso de selección se surte de la siguiente manera:

- Una muestra s_I de escuelas es seleccionada de U_I de acuerdo a un diseño de muestreo $p_I(s_I)$. El tamaño de la muestra de escuelas se denota como n_I. Nótese que S_I representa la muestra aleatoria de escuelas tal que $Pr(S_I = s_I) = p_I(s_I)$.
- Para cada escuela U_i $i = 1, \ldots, n_I$ seleccionada en la muestra s_I, se selecciona una muestra s_i de estudiantes de acuerdo a un diseño de muestreo $p_i(s_i)$. Nótese que S_i representa la muestra aleatoria de elementos tal que $Pr(S_i = s_i) = p_i(s_i)$.

NOTA 1. *Esta propuesta sugiere una variación frente al diseño de muestreo del TXXXXXXX. Mientras que esta propuesta sugiere la selección directa de estudiantes dentro de las escuelas, el diseño de muestreo del TXXXXXXX no seleccionaba alumnos dentro de la escuela, sino que seleccionaba aulas y luego todos los alumnos pertenecientes a esa aula específica eran seleccionados.*

NOTA 2. *Al considerar el tamaño del aula, es posible encontrar que si se muestrean muchas escuelas pequeñas esto puede ocasionar una reducción en la muestra total de estudiantes y conducir a estimaciones poco precisas y confiables. Para mitigar este problema, debería definirse un número mínimo de estudiantes por aula, que sería aceptable para cada país. En caso de que una escuela tenga aulas que no cumplan el mínimo requerido, es posible combinarla con otra para propósitos del muestreo.*

NOTA 3. *En el diseño de muestreo del TXXXXXXX, se seleccionaba (en la segunda etapa) un aula dentro de la escuela y todos los estudiantes que pertenecían a esa aula eran evaluados. La selección*

de estudiantes se realiza de esta manera, no considera que la variabilidad en los resultados entre aulas puede ser alta al suponer que dentro de la escuela es común particionar a los estudiantes con respecto a su desempeño académico. Por lo anterior, la selección de aulas, puede aumentar el error de muestreo y disminuir la precisión y confiabilidad de las estimaciones.

NOTA 3. *El esquema utilizado en TXXXXXXX es un caso particular de un muestreo de conglomerados de tamaños desiguales. Este tipo de muestreo es poco eficiente y tiende a aumentar el error de muestreo*[1] *de forma significativa. Por esta razón nuestra propuesta selecciona un número promedio de estudiantes dentro de cada escuela, induciendo un esquema de selección auto ponderado y más eficiente.*

4.2 Diseño de muestreo en cada etapa

Como ya se mencionó anteriormente, el diseño de muestreo para XXXXXXX seguiría dos etapas: una en la que se seleccionan las escuelas y otra en la que se escogen alumnos dentro de las aulas.

En la primera etapa se identifican todas las escuelas que hacen parte de la población objetivo (generación del marco muestral, siguiendo los criterios definidos en la sección 2); se separan en grupos mutuamente excluyentes, según las variables de estratificación explícita previamente definidas; y se seleccionan aquellas que harán parte del estudio, teniendo en cuenta un muestreo sistemático donde la probabilidad que tiene cada escuela de pertenecer a la muestra está determinada por el número de estudiantes en los grados de interés (medida de tamaño).

En esta etapa es importante tener en cuenta dos puntos. Primero, se seleccionará un número mayor de escuelas en los estratos explícitos más grandes. Segundo, la medida de tamaño permite que las escuelas con mayor cantidad de estudiantes tengan una mayor probabilidad de ser muestreadas (en comparación con las escuelas más pequeñas). Sin embargo, esta diferencia en las probabilidades de selección se compensa en la segunda etapa de muestreo, debido a que cada alumno dentro de las escuelas tiene igual probabilidad de ser elegida.

Para la segunda etapa se requiere contar con un listado de todas los estudiantes dentro de todas las aulas que tengan estudiantes en los grados de interés, para cada escuela seleccionada en la muestra. De forma aleatoria y teniendo en cuenta la misma probabilidad de selección, se elige una muestra de estudiantes.

4.2.1 Diseño de muestreo en las escuelas

Las escuelas se seleccionan en la primera etapa de muestreo, siguiendo un muestreo proporcional a la medida de tamaño previamente definida. Esta medida de tamaño (MOS) corresponde a la matrícula (número de estudiantes matriculados) en terceroo y sexto y es la variable que permite la selección

[1] ver Särndal, Swensson & Wretman (2003), pg. 133

de escuelas con probablidades desiguales, induciendo que las escuelas con mayor cantidad de estudiantes tengan una mayor probabilidad de ser seleccionadas en la muestra de la primera etapa.

La selección de escuelas se realizará de forma independiente a través de un muestreo sistemático proporcional[2] (**piPS**) sin reemplazo. Por otra parte, el algoritmo de Sunter[3] se encuentra detallado en Gutiérrez (2015) y Särndal, Swensson & Wretman (2003) y se implementará en este estudio utilizando la función `S.piPS` del paquete `TeachingSampling` de 'R. Särndal, Swensson & Wretman (2003) aseguran que la utilización de este diseño en la primera etapa contribuye a la reducción de varianza de los estimadores para los parámetros de interés (totales, razones, promedios y proporciones).

Una vez definida la medida de tamaño, las escuelas se clasifican según las variables de estratificación implícita dentro de cada estrato explícito. Primero se usa la primera variable de estratificación implícita, luego la segunda (dentro de los niveles generados por la primera) y así sucesivamente hasta aplicar todas las variables. La última variable de ordenamiento siempre será la medida de tamaño.

Luego de haber definido la medida de tamaño, se realiza un muestreo (sistemático) proporcional para seleccionar la muestra de escuelas. Para este proceso es importante contar con un marco de muestreo completo (donde cada escuela tenga su medida de tamaño definida) y con el conteo de escuelas a muestrear en cada estrato. El cociente entre estos dos valores: la medida de tamaño dividida entre el número de escuelas a muestrear en el estrato corresponde al intervalo de muestreo.

Por ejemplo, PISA usa esta medida para seleccionar las escuelas[4]: primero, las escuelas grandes, cuya medida de tamaño supere dicho intervalo son seleccionadas con probabilidad de inclusión forzosa; luego, se vuelve a calcular el intervalo de muestreo con una escuela menos en el denominador y se seleccionan nuevamente con certeza las escuelas con la medida de tamaño que supere ese valor. Este proceso se repite hasta que no se encuentren más escuelas que superen el intervalo establecido.

Posteriormente, se genera un número aleatorio entre cero y uno para cada estrato, el cual se utiliza para calcular un número de selección para cada escuela. El primer número de selección se obtiene multiplicando el número aleatorio por el intervalo de muestreo (este número identifica la primera escuela muestreada en el estrato). Al número obtenido se le suma el intervalo de muestreo para identificar la segunda escuela muestreada. El proceso se repite sumando el intervalo al último número obtenido e identificando así las escuelas muestreadas.

4.2.2 Diseño de muestreo dentro de las escuelas

La segunda etapa de muestreo corresponde la selección de los estudiantes que asisten a las escuelas seleccionadas en la primera etapa de muestreo y que deberán presentar la prueba, mediante los instrumentos cognitivos, además de los cuestionarios de factores asociados. Esta selección se realizará a través de un muestreo aleatorio simple sin reemplazo (**MAS**) dentro de cada estrato explícito.

[2] Cada escuela tendrá una probabilidad de inclusión desigual y proporcional al número de estudiantes en los grados terceroo y sexto

[3] Que es un caso particular de un algoritmo *sistemático proporcional* cuando se hace un ordenamiento conveniente por las variables de estratificación explícita, implícita y por la medida de tamaño.

[4] ver PISA (2012), Technical Report, pg. 74

Dentro de cada escuela seleccionada, la muestra de estudiantes (USM) se obtiene empleando el algoritmo de Fan-Muller & Rezucha (algoritmo para selección de unidades en un MAS) dentro de los estratos definidos en cada uno de los grupos de interés. Este algoritmo se encuentra descrito en detalle en Gutiérrez (2015) y puede ser implementado en diferentes funciones de `R` como por ejemplo la función `S.STSI` del paquete `TeachingSampling` (Gutiérrez, 2017).

4.3 Definición de los algoritmos de selección

El acceso y observación de los elementos de la población se establece mediante un algoritmo de muestreo, que es un mecanismo que asocia los elementos de la población con las unidades de muestreo definidas en el diseño.

4.3.1 Algoritmo de Sunter

En la primera etapa se seleccionan los municipios mediante un muestreo PPS utilizando el algoritmo de Sunter. Este algoritmo se encuentra detallado en (Gutiérrez, 2015. pág. 155) y es implementado en la función `S.piPS` del paquete `TeachingSampling`. Este es un procedimiento secuencial que funciona cuando los elementos de la población son ordenados descendentemente y cuando los elementos con valores más pequeños comparten las mismas probabilidades de inclusión. Este método asume la existencia de una medida de tamaño (MOS), notada como x, y consiste en:

1. Ordenar descendentemente la población de acuerdo con los valores que toma la medida de tamaño x_k.
2. Calcular $\pi_k = \frac{n*x_k}{\sum_{k\in U_i} x_k}$ para cada estudiante k pertenenciente a la escuela U_i.
3. Generar un número aleatorio $\xi_k \sim U(0,1)$.
4. Para $k = 1$, el primer elemento de la lista ordenada es incluido en la muestra sí y solamente sí $\xi_1 < \pi_1$.
5. Para $k \geq 2$, el k-ésimo elemento de la lista ordenada es incluido en la muestra sí y solamente sí

$$\xi_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i}\pi_k$$

donde n_{k-1} representa el número de elementos que ya han sido seleccionados al final del paso $k-1$.

4.3.2 Algoritmo de Fan-Muller-Rezucha

Al interior de cada uno de las escuelas se seleccionan a los estudiantes en cada estrato mediante el algoritmo de Fan-Muller-Rezucha implementado en el paquete `TeachingSampling` en la función 'S.STSI. El algoritmo está descrito detalladamente en Gutiérrez (2015), y consiste en recorrer el marco

de muestreo, elemento por elemento, y decidir la pertenencia o el rechazo del estudiante en la muestra. En general se supone que el marco de muestreo tiene N individuos, y se quiere seleccionar una muestra aleatoria de n individuos. Así, para el individuo k $(k = 1, 2, ..., N)$, se tiene que

1. Generar un número aleatorio $\xi_k \sim U(0, 1)$
2. Calcular
$$c_k = \frac{n - n_k}{N - k + 1}$$
donde n_k es la cantidad de objetos seleccionados en los $k - 1$ ensayos anteriores.
3. Si $\xi_k < c_k$, entonces el elemento k pertenece a la muestra.
4. Detener el proceso cuando $n = n_k$.

Dado que este algoritmo se detiene cuando $n = n_k$, resulta muy eficiente porque asegura una muestra aleatoria simple y en algunas ocasiones no se requiere recorrer todo el marco de muestreo.

4.4 Estratificación

El proceso de estratificación consiste en dividir las escuelas de la población objetivo en grupos o estratos que comparten las mismas características. Como se explicó en la sección 2, la estratificación puede ser explícita o implícita, y es una característica importante de esta propuesta porque hace más eficiente el diseño muestral y asegura la representatividad proporcional de los grupos de interés en la muestra.

En el XXXXXXX (manteniendo los criterios del TXXXXXXX) las variables de estratificación serán:

- Dependencia administrativa de la escuela: aquí encontramos escuelas con administración pública o privada.
- Área: en donde encontramos escuelas rurales y urbanas (según la definición de cada país).
- Grados en la escuela: Aquí encontramos escuelas que tienen solo terceroo, o solo sexto, o ambos grados.

La combinación de estas variables genera doce estratos, cuyo tamaño estará definido de acuerdo a la distribución de escuelas en cada país.

4.5 Selección de los reemplazos

El escenario ideal en la aplicación de cualquier prueba internacional es que todas las escuelas que son muestreadas originalmente accedan a ser parte del estudio; sin embargo, no es usual tener una tasa de participación del 100%. Así, para evitar una reducción en el tamaño de la muestra y garantizar un nivel de precisión y confiabilidad adecuados para el análisis estadístico, cada escuela cuenta con dos reemplazos en caso de que se rehúse a participar.

Estos corresponden, por lo general, a las escuelas que se encuentran inmediatamente antes y después de la originalmente muestreada en el marco muestral (que debe estar estrictamente ordenado por las variables de estratificación explícitas, implícitas y por la medida de tamaño descentemente). Estos reemplazos siempre se encontrarán en el mismo estrato explícito (aunque es posible que no haga parte del mismo estrato implícito) y tendrá un tamaño similar a la muestreada originalmente. Sin embargo, aun cuando las características son similares realizar estos cambios en la muestra puede producir sesgo, por esta razón es importante que los países se esfuXXXXXXXn por asegurar la participación de las escuelas originalmente muestreadas y solo usen los reemplazos en caso de que sea estrictamente necesario.

En países pequeños, donde puede ser difícil encontrar dos reemplazos para cada escuela, puede evaluarse la posibilidad de que una escuela sea un reemplazo potencial para dos escuelas originalmente muestreadas (sería un reemplazo real solo para una de ellas). También, si es elegida una escuela que se encuentra al inicio o al final de la lista, es posible seleccionar dos reemplazos consecutivos (después de ella, en caso de estar al inicio, y antes de ella, en caso de estar al final). Las siguientes gráficas muestran de forma explicativa el proceso de selección de los reemplazos de las escuelas seleccionadas.

Estratificación		Código escuela	Número estudiantes	Estado - Diseño
Explícita	Implícita			
Rural	Escuelas grandes	RG001	100	No seleccionada
		RG002	98	Reemplazo
		RG003	96	Seleccionada
		RG004	93	Reemplazo
		RG005	91	No seleccionada
		.	.	
		.	.	
		.	.	
	Escuelas medianas	RM001	70	No seleccionada
		RM002	67	Reemplazo
		RM003	65	Seleccionada
		RM004	62	Reemplazo
		RM005	61	No seleccionada
		.	.	
		.	.	
		.	.	
	Escuelas pequeñas	RP001	40	No seleccionada
		RP002	37	Reemplazo
		RP003	35	Seleccionada
		RP004	33	Reemplazo
		RP005	31	No seleccionada
		.	.	
		.	.	
		.	.	

Figura 4.1: Ejemplo de un esquema de selección de reemplazos en el estrato rural

Estratificación		Código escuela	Número estudiantes	Estado - Diseño
Explícita	Implícita			
Urbana	Escuelas grandes	UG001	160	No seleccionada
		UG002	155	Reemplazo
		UG003	153	Seleccionada
		UG004	152	Reemplazo
		UG005	155	No seleccionada
		.	.	
		.	.	
		.	.	
	Escuelas medianas	UM001	100	No seleccionada
		UM002	97	Reemplazo
		UM003	93	Seleccionada
		UM004	91	Reemplazo
		UM005	90	No seleccionada
		.	.	
		.	.	
		.	.	
	Escuelas pequeñas	UP001	40	No seleccionada
		UP002	37	Reemplazo
		UP003	35	Seleccionada
		UP004	33	Reemplazo
		UP005	31	No seleccionada
		.	.	
		.	.	
		.	.	

Figura 4.2: Ejemplo de un esquema de selección de reemplazos en el estrato urbano

Capítulo 5

Tamaño de muestra

El diseño muestral propuesto para el XXXXXXX sigue un método de selección por etapas. Los estudiantes en una misma escuela y una misma aula tienden a compartir distintas características, las cuales están relacionadas con los logros de aprendizaje. Esto hace que la variabilidad observada en los resultados *dentro* de las escuelas sea menor a la observada *entre* ellas, lo cual puede arrojar una estimación precisa o no, dependiendo del tamaño de la muestra.

Dado un diseño muestral complejo, si el tamaño de la muestra es muy pequeño, los resultados podrían no ser representativos del país evaluado y podrían indicar niveles de variabilidad poco confiables. Para evitar este problema, los tamaños de muestra se calcularon haciendo uso de la metodología del *efecto diseño de Kish*, la cual corrige el tamaño de muestra calculado bajo un muestreo aleatorio simple de estudiantes.

Los tamaños de muestra que se obtienen después de aplicar este factor de ajuste pueden variar en cada país, debido a procesos de sobremuestreo o a ajustes basados en el error muestral estimado de los estratos. Sin embargo, se definirá un número mínimo de escuelas que deben ser evaluadas en cada país participante y, en caso de que un país tenga un menor número de escuelas elegibles, todas deberán participar en el estudio.

5.1 Precisión de la inferencia

En el XXXXXXX la precisión de la inferencia sobre el logro de aprendizaje de los estudiantes es primordial. Todas las tareas desde la definición de la población objetivo, incluyendo especialmente la elaboración del marco muestral y la identificación de variables de estratificación implícita y explícita apoyan este objetivo. Para que la prueba cumpla con este propósito se deben alcanzar tasas de participación satisfactorias, tener en cuenta la escala de la prueba con media 700 y desviación estándar 100, y además cumplir algunos criterios estadísticos.

En TIMSS y PIRSLS, por ejemplo, el error estándar no debe ser mayor a 0,035 desviaciones estándar

para el puntaje promedio del país. Con una desviación estándar de 100, esto corresponde a un intervalo de confianza del 95% de más y menos siete puntos (margen de error) para el promedio. Además, las estimaciones de la muestra de cualquier estimación de porcentaje de estudiantes deben tener un intervalo de confianza de más y menos 3.5%.

La precisión esperada para este estudio piloto estará dada por un margen de error δ de 0.035 desviaciones estándar sobre la escala de la prueba.

5.2 Tamaño de muestra por país

El proceso de definición del tamaño de la muestra en cada país sigue los siguientes pasos:

1. Se obtiene el tamaño de muestra de estudiantes por país, suponiendo un muestreo aleatorio simple.
2. Se calcula el tamaño de muestra de estudiantes por país corregido por el efecto de diseño *DEFF*.
3. Se calcula el tamaño de muestra de las escuelas.
4. Se distribuye el tamaño calculado entre los estratos explícitos (estratos más grandes tendrán asociado un mayor tamaño de muestra).
5. Se calculan los errores muestrales para cada estrato y se realizan los ajustes pertinentes.
6. Se seleccionan las escuelas dentro de cada estrato y a los estudiantes por grado de interés dentro de cada escuela.

Nótese que la muestra de estudiantes se obtiene haciendo un submuestreo de los estudiantes por grado de interés en cada escuela. Es decir, el muestreo se realiza entre grupos de estudiantes de una misma escuela, que tienen la misma probabilidad de ser elegidos. Este método facilita la selección de la muestra y presenta errores de muestreo menores a los que se obtendrían muestreando directamente aulas dentro de la escuela.

5.2.1 Tamaño de muestra para estimar un puntaje promedio

La estrategia metodológica que se utilizará en esta propuesta consta de tres pasos y está alineada con propuestas de pruebas algunas otras pruebas estandarizadas a nivel internacional (Foy and Joy, 2015):

1. Identificar el tamaño de muestra suponiendo un muestreo aleatorio simple sin reemplazo (MAS), de acuerdo a la siguiente expresión:

$$n_{MAS} = \frac{z_\alpha^2 S_y^2}{\delta^2 + \frac{z_\alpha^2 S_y^2}{N}}$$

En donde z_α^2 representa el pXXXXXXXntil asociado a una densidad normal estándar que cubre el $(1-\alpha)\times$ 100% de la distribución, S_y^2 es la varianza poblacional de la variable de interés (y), δ es el máximo error admisible en la escala, N es el número de estudiantes del grado en el país.

2. Calcular el efecto de diseño asociado a cada país, como función del coeficiente de correlación[1] intraclase ρ y del número de estudiantes m que serán seleccionados en promedio en cada escuela:

$$DEFF(\rho, m) = 1 + (m-1)\rho$$

3. Calcular el tamaño de muestra final para un muestreo en dos etapas realizando la multiplicación de las anteriores expresiones:

$$n_{2E} = n_{MAS} DEFF(\rho, m)$$

4. Calcular el número de escuelas que se deben seleccionar en la primera etapa de la siguiente manera:

$$n_I \approx \frac{n_{2E}}{m}$$

Por ejemplo, en un país con $N = 100000$ estudiantes con coeficiente de correlación intraclase $\rho = 0.22$ y con desviación estándar de 90 puntos en la prueba, el tamaño de muestra que garantiza un error de máximo $\delta = 10$ puntos en la prueba, con una confianza estadística del 95% ($z_\alpha = 1.96$), entonces

$$n_{MAS} = \frac{1.96^2 * 90^2}{10^2 + \frac{1.96^2 * 90^2}{100000}} = 310$$

Luego, suponiendo una selección promedio de 35 estudiantes por escuela, el tamaño de muestra final será

$$n_{2E} = 310 * [1 + (35-1)0.22] = 2629$$

Estos 2629 estudiantes serán seleccionados en $n_I = \frac{2629}{35} \approx 75$ escuelas dentro del país.

5.2.2 Tamaño de muestra para estimar el porcentaje de respuestas correctas

Por otro lado, en el XXXXXXX también es posible considerar como parámetro de interés el porcentaje de estudiantes que responde acertadamente un ítem. Para calcular el tamaño de muestra necesario

[1] Esta medida toma valores positivos si los estudiantes dentro de las escuelas tienen un desempeño similar y ligeramente negativo cuando el logro de los estudiantes dentro de las escuelas es muy disperso. Es decir, el coeficiente informa qué tan similares son los estudiantes dentro de las escuelas, proporcionando una medida de homogeneidad dentro de las escuelas.

para estimar esta proporción con un error máximo admisible de $\delta\times$ 100%, es necesario recurrir a la siguiente expresión para calcular el tamaño de muestra en un MAS

$$n_{MAS} = \frac{z_\alpha^2 P(1-P)}{\delta^2 + \frac{z_\alpha^2 P(1-P)}{N}}$$

En donde P representa la proporción de estudiantes que acierta el ítem, la cual como se supone desconocida se fijará en $P = 0.5$, pues este valor hace máxima la varianza del estimador y por tanto nunca inducirá una subestimación del tamaño de muestra.

Siguiendo con el ejemplo anterior, para estimar la proporción de estudiantes que aciertan una pregunta, admitiendo un error máximo de $\delta = 5\%$, se tendría que:

$$n_{MAS} = \frac{1.96^2 * 0.5(1-0.5)}{0.05^2 + \frac{1.96^2 * 0.5(1-0.5)}{100000}} = 382$$

Luego, suponiendo una selección promedio de 35 estudiantes por escuela, el tamaño de muestra final será

$$n_{2E} = 382 * [1 + (35-1)0.22] = 3239$$

Estos 3239 estudiantes serían seleccionados en $n_I = \frac{3239}{35} \approx 93$ escuelas dentro del país.

5.3 Coeficientes de correlación intraclase en el TXXXXXXX

Nótese que un insumo fundamental para calcular los tamaños de muestra son los coeficientes de correlación intraclase. Debido a que UNESCO-OREALC ha provisto las bases de datos de los anteriores estudios en la web, es posible calcular estos coeficientes para cada país. La siguiente tabla muestra los coeficientes de correlación intraclase para las pruebas del TXXXXXXX para grado terceroo (lectura, matemáticas) y para grado sexto (lectura, matemáticas y ciencias).

País	terceroo Lectura	terceroo Matemáticas	Sexto Lectura	Sexto Matemáticas	Sexto Ciencias
Argentina	0.23	0.26	0.21	0.27	0.22
Brasil	0.25	0.34	0.21	0.29	0.21
Chile	0.17	0.22	0.15	0.24	0.24
Colombia	0.31	0.36	0.30	0.32	0.23
Costa Rica	0.17	0.19	0.17	0.20	0.18
Ecuador	0.24	0.27	0.28	0.30	0.25
Guatemala	0.30	0.30	0.20	0.21	0.21
Honduras	0.28	0.35	0.30	0.32	0.25

País	terceroo Lectura	terceroo Matemáticas	Sexto Lectura	Sexto Matemáticas	Sexto Ciencias
México	0.24	0.23	0.27	0.22	0.24
Nicaragua	0.24	0.28	0.20	0.21	0.22
Panamá	0.28	0.37	0.29	0.33	0.28
Paraguay	0.29	0.38	0.33	0.32	0.30
Perú	0.39	0.45	0.42	0.40	0.35
República Dominicana	0.19	0.26	0.19	0.12	0.14
Uruguay	0.23	0.23	0.17	0.21	0.17

5.4 Tablas de muestreo

Las tablas de muestreo son una herramienta que permite decidir acerca del tamaño de muestra con relación a los costos de la implementación del estudio. Estas tablas están basadas en un diseño de muestreo en dos etapas estratificado. En la primera etapa, se va a seleccionar las escuelas con un algoritmo que le da mayor probabilidad de inclusión a las escuelas más grandes y que permite seleccionar los reemplazos de acuerdo a la estratificación implícita. En la segunda etapa, se seleccionan estudiantes dentro de cada escuela seleccionada.

A modo de ejemplo, tomamos los datos del TXXXXXXX para los países participantes[2] y calculamos el coeficiente de correlación intraclase para todos los países[3] y se generaron distintos escenarios de tamaños de muestra de escuelas, de estudiantes dentro de la escuela, y por ende de estudiantes en el país.

Todos estos escenarios son calculados para estimar la proporción de respuestas correctas en los ítems de pilotaje. Estas estimaciones finales tendrán un error máximo de 7.5% puntos porcentuales. Note que todos los escenarios inducen el mismo error de muestreo y la única diferencia radica en los costos de aplicación de la prueba. Todas las tablas fueron generadas utilizando las funciones `ICC`y `ss2s4p` del paquete `samplesize4surveys` (Gutiérrez, 2017) del software estadístico `R`.

5.4.1 Primer escenario ($m = 10$)

La siguiente tabla propone un tamaño de muestra (tanto de las escuelas como de los estudiantes dentro de cada escuela con una confianza del 95% y un error máximo admisible de 7.5%) teniendo

[2]Para los países que no participaron en el TXXXXXXX se asumió que el coeficiente de correlación intraclase es igual al promedio de los demás países. Note que se podría indagar más y estimar este coeficiente con algún país que tenga un comportamiento similar.

[3]Este cálculo se realizó para la prueba de matemáticas de grado sexto.

en cuenta la selección de 10 alumnos en promedio en cada escuela incluida en la muestra de la primera etapa para cada país.

Países	rho	DEFF(m=10)	Muestras de escuelas	Muestra de estudiantes
Argentina	0.27	3.42	59	585
Brasil	0.29	3.62	62	619
Chile	0.24	3.12	53	534
Colombia	0.32	3.9	67	667
Costa Rica	0.20	2.83	48	484
Ecuador	0.30	3.72	64	636
Guatemala	0.21	2.85	49	487
Honduras	0.32	3.88	66	663
México	0.22	2.99	51	511
Nicaragua	0.21	2.92	50	499
Panamá	0.33	4.01	69	686
Paraguay	0.32	3.87	66	662
Perú	0.40	4.63	79	792
República Dominicana	0.12	2.1	36	359
Uruguay	0.21	2.87	49	491
Venezuela	0.26	3.38	58	578
Cuba	0.26	3.38	58	578
El Salvador	0.26	3.38	130	1298
Bolivia	0.26	3.38	130	1298

5.4.2 Segundo escenario ($m = 15$)

La siguiente tabla propone un tamaño de muestra (tanto de las escuelas como de los estudiantes dentro de cada escuela con una confianza del 95% y un error máximo admisible de 7.5%) teniendo en cuenta la selección de 15 alumnos en promedio en cada escuela incluida en la muestra de la primera etapa para cada país.

Países	rho	DEFF(m=15)	Muestras de escuelas	Muestra de estudiantes
Argentina	0.27	4.77	54	816
Brasil	0.29	5.07	58	867
Chile	0.24	4.31	49	737
Colombia	0.32	5.51	63	942
Costa Rica	0.20	3.84	44	657
Ecuador	0.30	5.23	60	894
Guatemala	0.21	3.88	44	663
Honduras	0.32	5.47	62	935
México	0.22	4.09	47	699

Países	rho	DEFF(m=15)	Muestras de escuelas	Muestra de estudiantes
Nicaragua	0.21	3.98	45	681
Panamá	0.33	5.68	65	971
Paraguay	0.32	5.46	62	934
Perú	0.40	6.64	76	1135
República Dominicana	0.12	2.71	31	463
Uruguay	0.21	3.91	45	669
Venezuela	0.26	4.7	54	804
Cuba	0.26	4.7	54	804
El Salvador	0.26	4.7	54	804
Bolivia	0.26	4.7	54	804

5.4.3 tercero escenario ($m = 20$)

La siguiente tabla propone un tamaño de muestra (tanto de las escuelas como de los estudiantes dentro de cada escuela con una confianza del 95% y un error máximo admisible de 7.5%) teniendo en cuenta la selección de 20 alumnos en promedio en cada escuela incluida en la muestra de la primera etapa para cada país.

Países	rho	DEFF(m=20)	Muestras de escuelas	Muestra de estudiantes
Argentina	0.27	6.11	52	1045
Brasil	0.29	6.53	56	1117
Chile	0.24	5.49	47	939
Colombia	0.32	7.12	61	1218
Costa Rica	0.20	4.86	42	831
Ecuador	0.30	6.74	58	1153
Guatemala	0.21	4.91	42	840
Honduras	0.32	7.07	60	1209
México	0.22	5.2	44	889
Nicaragua	0.21	5.04	43	862
Panamá	0.33	7.35	63	1257
Paraguay	0.32	7.06	60	1207
Perú	0.40	8.66	74	1481
República Dominicana	0.12	3.32	28	568
Uruguay	0.21	4.95	42	846
Venezuela	0.26	6.03	52	1031
Cuba	0.26	6.03	52	1031
El Salvador	0.26	6.03	52	1031
Bolivia	0.26	6.03	52	1031

NOTA: *A modo de referencia, los participantes en el estudio piloto[4] del TXXXXXXX fueron 15.484 estudiantes en tercero grado y 15.840 en sexto grado. Para un total de 31.324 niños en 16 economías (15 países y el Estado mexicano de Nuevo León). Por lo tanto, TXXXXXXX tuvo un promedio por país de 968 estudiantes en tercero grado y de 990 estudiantes en sexto grado.*

NOTA: *Otro enfoque común que se utiliza para calcular el tamaño de muestra requerido en el estudio piloto es satisfacer una muestra mínima de estudiantes presentando cuadernillos con ítems de pilotaje. Por ejemplo, la muestra del piloto[5] para TIMSS y PIRLS se extrae al mismo tiempo y de la misma población de escuelas que la muestra completa. El requisito de tamaño de muestra de la prueba piloto es de 200 estudiantes por cuadernillo. El tamaño final de la muestra del piloto es una función del número de cuadernillos cognitivos que se están probando en el campo. Normalmente, PIRLS tiene cuatro cuadernillos de prueba y por lo tanto requiere una muestra piloto de no menos de 800 estudiantes, mientras que TIMSS con seis cuadernillos requiere una muestra de más de 1200 estudiantes en cada grado.*

La siguiente tabla muestra la distribución de escuelas en el piloto del TXXXXXXX para algunas economías participantes.

País	terceroo	Sexto
Argentina	46	43
Brasil	34	27
Chile	37	35
Colombia	54	34
Guatemala	55	49
Honduras	52	50
México	48	48
Nicaragua	73	54
Nuevo León	42	42
Paraguay	55	55
Perú	54	49
República Dominicana	54	50
Uruguay	43	45

5.5 Tratamiento de las escuelas pequeñas

Tener muchas escuelas pequeñas en la muestra final es poco deseable, puesto que puede reducir significativamente el tamaño de la muestra de estudiantes esperada para el país. Se considera que una escuela es pequeña cuando el número de estudiantes elegibles es menor al número de estudian-

[4] ver UNESCO – OREALC (2016), Reporte Técnico TXXXXXXX pg. 66

[5] ver Joncas, M. & Foy, P. (2013), pg. 8

tes que se esperaba seleccionar dentro de ella. Por ejemplo, en PISA[6] se definieron las siguientes categorías, teniendo en cuenta que se esperaba seleccionar 35 estudiantes en cada escuela:

- Escuelas grandes: el número de estudiantes elegibles era mayor a 35.
- Escuelas medianas: el número de estudiantes elegibles estaba entre 18 (la mitad del valor que se esperaba seleccionar) y 35.
- Escuelas pequeñas: el número de estudiantes elegibles era mayor a dos, pero menor a 18.
- Escuelas muy pequeñas: el número de estudiantes elegibles era cero, uno o dos.

Para seleccionar un número apropiado de escuelas pequeñas en la muestra, estas se submuestrean, aumentando proporcionalmente el número de grandes escuelas. Las escuelas pequeñas se submuestrean con un factor de dos (su probabilidad de inclusión se reduce a la mitad) y las escuelas muy pequeñas con un factor de cuatro (su probabilidad de inclusión se reduce tres cuartas partes). Para definir si un país debe realizar este proceso se consideran los siguientes casos:

- Si el porcentaje de estudiantes en escuelas pequeñas y muy pequeñas es igual o mayor al 1%, entonces estas se submuestrean y se aumenta el tamaño de la muestra para compensarlas.
- Si el porcentaje de estudiantes en escuelas pequeñas y muy pequeñas es menor al 1% y el porcentaje de estudiantes en escuelas medianas es igual o mayor al 4%, entonces no es necesario el proceso de submuestreo, pero el tamaño de la muestra se incrementa.
- Si ninguna de las anteriores condiciones se cumple, quiere decir que el número de estudiantes en las escuelas pequeñas es muy pequeño para afectar la muestra. Por lo tanto, no es necesario el proceso de submuestreo y tampoco el aumento de la muestra.

En caso de que el proceso de submuestreo sea necesario, se propone adaptar el siguiente algoritmo (asumiendo un tamaño inicial de muestra igual a n_I escuelas y n_{2E} estudiantes):

- Definir la proporción de estudiantes en cada una de las cuatro categorías de escuela definidas (la suma de ellas debe ser 1).
- Calcular la siguiente cantidad
$$L = 1 + \frac{3}{4}p_{mp} + \frac{1}{2}p_{pq}$$
en donde p_{mp} es la proporción de estudiantes en escuelas muy pequeñas y p_{pq} es la proporción de estudiantes en escuelas pequeñas.
- El tamaño mínimo de la muestra para las escuelas grandes será
$$n_I^{eg} = n_I * L * p_{eg}$$
en donde p_{eg} hacer referencia a la proporción de estudiantes en escuelas grandes. Este tamaño puede incrementarse según los requerimientos del país.
- Calcular el valor promedio de estudiantes elegibles en las escuelas moderadamente pequeñas, pequeñas y muy pequeñas. El tamaño de la muestra para las escuelas medianas será
$$n_I^{me} = \frac{n_{2E}}{m_{me}} * p_{me}$$

[6] ver PISA Technical Report, págs. 76 y 77

en donde m_{me} y p_{me} es el valor promedio de estudiantes elegibles en las escuelas medianas y la proporción de estudiantes en escuelas medianas, respectivamente.

- Definir el tamaño de la muestra para las escuelas pequeñas como

$$n_I^{pq} = \frac{n_{2E}}{2 * m_{me}} * p_{ep} * L$$

- Definir el tamaño de la muestra para las escuelas muy pequeñas como

$$n_I^{mp} = \frac{n_{2E}}{4 * m_{mp}} * p_{mp} * L$$

en donde m_{mp} y p_{mp} es el valor promedio de estudiantes elegibles en las escuelas muy pequeñas y la proporción de estudiantes en escuelas muy pequeñas, respectivamente.

- Definir el tamaño final de la muestra, sumando los tamaños obtenidos para cada categoría de escuela.

Es posible realizar este análisis de escuelas pequeñas para los estratos y sobremuetras de interés, indicando cuántas escuelas de cada categoría deben ser seleccionadas en cada estrato explícito (definiendo previamente si era necesario realizar submuestreo).

5.6 Tamaño de muestra ajustado por ausencia de respuesta

La tasa de participación de las escuelas y la tasa de respuesta de los estudiantes pueden afectar los análisis estadísticos si son muy bajas. Cada país realiza un esfuerzo para que todas las escuelas seleccionadas en la muestra accedan a participar y para que los estudiantes elegidos dentro de ellas contesten la prueba. Sin embargo, en ocasiones es difícil obtener el compromiso de las escuelas y, más aún el de los estudiantes, quienes pueden no asistir el día de la evaluación o asistir pero dejar el cuadernillo en blanco.

Para mitigar esta situación es necesario aumentar el tamaño de muestra, así, en caso de que se pierdan estudiantes o escuelas, esto no afectará la representatividad de la muestra. De esta forma, siendo ϕ la probabilidad esperada de que una escuela acceda a aplicar la prueba, entonces el tamaño de muestra de escuelas, ajustado por este factor será:

$$n_I^{(adj)} = \frac{n_I}{\phi}$$

Por ejemplo, considerando una tasa de respuesta esperada del 85% y una muestra mínima de 50 escuelas, el incremento por ausencia de respuesta sería de 9 escuelas aproximadamente, así el tamaño de muestra ajustado sería $n_I^{(adj)} = \frac{50}{0.85} = 59$ escuelas.

5.6.1 Afijación de la muestra en los estratos

En el diseño de muestreo propuesto se realizará el cálculo del tamaño de muestra global y, a partir de este, se realizará la asignación de los tamaños de muestra dentro de los estratos con el fin de garantizar suficiente información dentro de estos, para luego cumplir los objetivos del presente estudio.

Como el objetivo principal del XXXXXXX es comparar los resultados de los logros de aprendizaje en cada uno de los sistemas educativos para cada uno de los países participantes, se debe calcular un tamaño de muestra de forma independiente para cada estrato y así garantizar la precisión determinada en cada uno de ellos.

Si N_{Ih} es el total de escuelas en el estrato h para uno de los países del estudio (con H, el total de estratos definidos para este), la asignación de la muestra de escuelas en cada estrato se realiza por medio de una afijación de potencia con $0.5 \leq \alpha \leq 1$ empleando la siguiente fórmula:

$$n_{Ih} = \frac{N_{Ih}^{\alpha}}{\sum_{h=1}^{H} N_{Ih}^{\alpha}} * n_I$$

Con n_I el número total de escuelas en la muestra para el país en cuestión.

NOTA 1: *si $\alpha = 1$, entonces el método de afijación de potencia induce una repartición proporcional al tamaño de cada estrato.*

NOTA 2: *si $\alpha < 1$, entonces el método de afijación de potencia induce un mayor tamaño de muestra en aquellos estratos que tienen una proporción de estudiantes menor que otros estratos más grandes.*

Capítulo 6

Estrategia de estimación

6.1 Estimadores de los parámetros de interés

Con base en los parámetros definidos anteriormente, y teniendo en cuenta el diseño de muestreo que se propone en este documento, se definen las siguientes expresiones que inducen una estructura inferencial insesgada.

- **Estimación de un total poblacional**: para estimar el total poblacional de una variable de interés se utiliza el estimador de Horvitz-Thompson dado por la siguiente ecuación (Särndal, Swensson & Wretman, 2003):

$$\hat{t}_{y,\pi} = \sum_{s} w_k^{(design)} y_k$$

En donde $w_k^{(design)} = 1/\pi_k$ es el inverso de la probabilidad de inclusión ($\pi_k = Pr(k \in s)$) del k-ésimo estudiante en la muestra. El estimador anterior puede ser reescrito considerando el diseño muestral propuesto como se especifica en la siguiente expresión:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \sum_{i \in s_{hI}} \sum_{k \in s_i} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{i \in s_{hI}} \frac{1}{\pi_{Ihi}} \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{h=1}^{H} \sum_{i \in s_{hI}} \frac{t_{xh}}{nx_i} \sum_{k \in s_i} \frac{N_i}{n_i} y_k$$

En donde

- i denota el subíndice de las escuelas seleccionadas en la muestra estratificada de la primera etapa s_{hI},
- h denota el subíndice del estrato explícito y H denota el número total de estratos explícitos definidos por cada país,
- s_i hace referencia a la muestra de estudiantes en la escuela i,

- $\pi_{Ihi} \propto \frac{n_{Ih}x_i}{t_{xh}} \leq 1$ es la probabilidad de inclusión de la i-ésima escuela en el estrato explícito h,
- n_{Ih} es el número de escuelas seleccionadas en el estrato explicito h.
- x_i es la medida de tamaño de la i-ésima escuela en el estrato explícito h,
- t_{xh} es el total agregado de la medida de tamaño x en el estrato h,
- N_i es el número de estudiantes en la escuela i,
- n_i es el número de estudiantes seleccionados en la muestra de la escuela i.
- **Estimación de una razón poblacional**: para obtener la estimación de una razón ($\hat{R}$) se deben estimar los totales asociados a una variable de interés y, y el total asociado a una variable de interés $\mathbf{Z}$. Posteriormente realizar el cociente entre estas estimaciones, es decir:

$$\hat{R} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}} = \frac{\sum_s w_k^{(design)} y_k}{\sum_s w_k^{(design)} z_k}$$

Para la estimación de los anteriores totales se utilizan las mismas expresiones utilizadas para estimar un total poblacional.

- **Estimación del promedio poblacional**: para estimar un promedio basta con aplicar las expresiones para estimar una razón considerando $z_k = 1 \; \forall k \in U$; es decir, $\hat{t}_{z,\pi} = \hat{N}$. Por tanto, la estimación del promedio se estima como sigue:

$$\tilde{Y} = \frac{\hat{t}_{y,\pi}}{\hat{N}} = \frac{\sum_s w_k^{(design)} y_k}{\sum_s w_k^{(design)}}$$

Tanto $\hat{t}_{y,\pi}$ como $\hat{N}$ son estimaciones de totales. Para estimar estos totales son utilizadas las expresiones detalladas anteriormente para estimar un total poblacional. Es posible también estimar insesgadamente al dividir por N, aunque es posible que esta última estimación arroje una mayor varianza.

- **Estimación de una proporción poblacional**: para estimar la proporción poblacional se siguen los mismos principios que en la estimación de un promedio, pero sobre una variable dicotómica z_{dk} que toma el valor de 1 si el k-ésimo individuo tiene el atributo de interés d y de 0 en otro caso. Se obtiene la estimación de una proporción realizando la estimación del total de la variable y dividido sobre la estimación del total poblacional como se presenta a continuación:

$$\tilde{P} = \frac{\hat{t}_{z_d,\pi}}{\hat{N}} = \frac{\sum_s w_k^{(design)} z_{dk}}{\sum_s w_k^{(design)}}$$

6.2 Pesos de muestreo

Los pesos de muestreo se calculan para escuelas y estudiantes, según el diseño muestral del estudio para el piloto y aplicación principal del XXXXXXX. Como se mencionó en secciones anteriores, el muestreo se realiza en dos etapas. En cada una de las etapas, la unidad que será seleccionada, ya sean escuelas o estudiantes, tiene una probabilidad de ser elegida (en el caso de las escuelas dependiendo de la medida de tamaño o del número de estudiantes en la escuela). Los pesos de muestreo reflejan esas probabilidades de selección, teniendo en cuenta procesos de ajuste por ausencia de respuesta y el recorte de pesos en cada estrato (para reducir el error estándar al reducir la influencia de los sub-grupos pequeños de escuelas o estudiantes).

Estos pesos evidencian el diseño de la muestra y corresponden al recíproco de la probabilidad de inclusión de la unidad que será seleccionada (con algunos ajustes por las razones ya mencionadas). Dado el diseño bietápico, primero se calculan los pesos muestrales de las escuelas y después los del estudiante (considerando la probabilidad de ser elegido, dado que la escuela ya fue seleccionada en la muestra de la primera etapa. En principio, el peso muestral total de los estudiantes se define como el producto de estos dos resultados y es el que se requiere para el cálculo de estimadores de las características de la población objetivo, que sean imparciales o consistentes.

En el caso de las escuelas, aquellas más grandes tienen una mayor probabilidad de inclusión, en comparación con aquellas más pequeñas. Por lo tanto, los pesos muestrales para las escuelas más grandes tienden a ser más pequeños que los asociados a las escuelas más pequeñas. Además, es importante destacar que estos pesos indican el número de escuelas que la escuela seleccionada en la muestra está representando.

En el caso de los estudiantes, las probabilidades de inclusión reflejan la probabilidad de ser seleccionados dado que la escuela fue seleccionada en la muestra, es decir, se trata de una probabilidad condicional. Dentro de la escuela, cada estudiante tiene la misma probabilidad de ser elegido. Así, los pesos muestrales de los estudiantes en cada escuela serán los mismos.

Las muestras de escuelas y estudiantes deben ser representativas de la población objetivo y están diseñadas para obtener estimadores precisos, considerando un margen de error específico. Una vez se han recolectado todos los datos (provenientes de los instrumentos cognitivos y de los cuestionarios que contestan estudiantes, familias, profesores y directores), las estimaciones estadísticas se realizan teniendo en cuenta los pesos de muestreo como factor de ponderación, por esta razón es muy importante calcularlos correctamente.

Nótese que el peso muestral total no será el mismo para todos los estudiantes y esto puede deberse a varias razones. Por ejemplo, la información de los marcos de muestreo no fue exacta y una escuela que se esperaba fuera grande en realidad fue pequeña, lo que modifica la probabilidad de selección de los estudiantes dentro de ella (será mayor en comparación al resto de estudiantes de la muestra); o también puede suceder que la ausencia de respuesta en una escuela conduzca a una subrepresentación de los estudiantes en el estrato al que la escuela pertenece. En cualquier caso es necesario realizar un ajuste para que al final los pesos de muestreo representen correctamente a la población de interés.

6.2.1 Pesos de muestreo para las escuelas

El peso muestral de la escuela es el recíproco de su probabilidad de inclusión (proporcional a su medida de tamaño), con algunos ajustes como la corrección por ausencia de respuesta. Recordemos que la medida de tamaño para XXXXXXX son los estudiantes en los grados de interés. Para minimizar el efecto de las escuelas pequeñas en el cálculo de los pesos muestrales, es posible modificar la medida de tamaño. Por ejemplo, en PISA 2015 el tamaño esperado de la muestra por escuela era 35 estudiantes. Si el número de estudiantes de 15 años (medida de tamaño) es menor a ese valor esperado, pero mayor a 17, la medida de tamaño era 35. Si el número de estudiantes de 15 años era menor o igual a 17, pero mayor a 3, entonces la medida de tamaño es la mitad del tamaño esperado de la muestra por escuela. Y, si en la escuela había entre uno y dos estudiantes de 15 años, la medida de tamaño era la cuarta parte del tamaño esperado de la muestra por escuela.

El factor de expansión a nivel de la escuela (UPM) i en el estrato explícito h está dado por la siguiente expresión:

$$w_{hi}^{(school)} = \frac{1}{\pi_{Ihi}} \propto \frac{t_{xh}}{n_{Ih} x_i}$$

En donde π_{Ihi} se definió anteriormente y corresponde a la probabilidad de inclusión de primer orden de la escuela i y es obtenida a partir de la implementación del diseño muestral proporcional a la medida de tamaño.

6.2.2 Pesos de muestreo para los estudiantes

Las probabilidades de inclusión para los estudiantes indican la probabilidad de ser seleccionados dado que la escuela a la que pertenece fue seleccionada en la muestra, es decir, como se mencionó anteriormente, se trata de una probabilidad condicional. El peso de muestreo para los estudiantes es el recíproco de esa probabilidad, con dos ajustes: primero se aplica un factor de corrección por ausencia de respuesta y después se usa un recorte definido.

Cabe resaltar que todos los estudiantes de una misma escuela tienen igual probabilidad de ser elegidos. Si todos los estudiantes en un grado de interés (o ambos) son seleccionados, su peso de muestreo será igual a uno (puesto que cada estudiante se representa a si mismo), pero si no todos fueron seleccionados, los pesos de muestreo serán mayores.

En el caso de la segunda etapa, en donde el muestreo de los estudiantes es aleatorio simple sin reemplazo, el peso de muestreo de un estudiante k adscrito a la escuela i en el estrato explícito h está dada por la siguiente expresión:

$$w_{k|i}^{(student|school)} = \frac{1}{\pi_{k|i}} = \frac{N_i}{n_i}$$

Por la anterior, y debido a que el muestreo es independiente en cada etapa, el factor de expansión general para un estudiante que pertenece a una escuela del estrato explícito h está dado por la multiplicación de los anteriores pesos:

$$w_k^{(design)} = w_{hi}^{(school)} w_{k|i}^{(student|school)} \propto \frac{t_{xh}}{n_{Ih} x_i} \frac{N_i}{n_i}$$

6.2.3 Modificación de los pesos de estudiantes: *Student house-weight*

El método *house-weight* permite realizar transformaciones proporcionales en los pesos de muestreo para que sumen un tamaño determinado dentro de cada país. Estas transformaciones no afectan las estimaciones de medias o parámetros de interés, aunque sí modifican la suma del propio peso de muestreo.

El *house-weight* le da un peso diferente a cada país, dependiendo del tamaño de población. Se utiliza cuando el error estándar está siendo calculado en la suma de los pesos de muestreo y no en el propio tamaño de la muestra y se requiere corregir esta situación (con los avances en los softwares de análisis de datos, esta necesidad es cada vez menor).

Este tipo de ponderación, también llamado peso normalizado, se utiliza cuando los análisis son sensibles al tamaño de la muestra y representa, esencialmente, una transformación lineal del peso total del estudiante de modo que la suma de los pesos sea igual al tamaño de la muestra de cada país n. En general, se tiene que

$$w_k^{(house)} = \frac{w_k^{(design)}}{\sum_s w_k^{(design)}} * n$$

6.2.4 Modificación de los pesos de estudiantes: *Student senate-weight*

Al igual que la modificación anterior, el método *senate-weight* permite realizar transformaciones proporcionales en los pesos de muestreo para que sumen un tamaño determinado dentro de cada país, sin afectar las estimaciones de medias o coeficientes (se modifica la suma del propio peso de muestreo). En este método, a diferencia de *house-weight*, los países tienen el mismo peso, sin que se considere el tamaño de la población (las poblaciones objetivo dentro de cada país participante serán del mismo tamaño), por lo que es útil para realizar análisis entre países.

Este método se utiliza cuando los análisis involucran más de un país porque es el peso total del estudiante escalado de tal manera que los pesos de senado de todos los estudiantes suman a 500 (o 1000)[1] en cada país. De esta forma, la siguiente ecuación define la forma estructural de estos pesos:

[1] ver Joncas, M. & Foy, P. (2013)

$$w_k^{(senate)} = \frac{w_k^{(design)}}{\sum_s w_k^{(design)}} * \kappa$$

En donde κ es un número que se fija de antemano, y que puede ser igual a 500 o 1000.

6.2.5 Acerca de los pesos de muestreo de estudiantes de terceroo y sexto

Por la naturaleza del estudio, en donde un mismo marco de muestreo cuenta con escuelas de grados terceroo y sexto, se necesita plantear un diseño que contemple los pesos de muestreo que sean representativos para esta realidad. Para empezar, utilizaremos tres estratos, definido a continuación.

- Estrato 1: escuelas que cuentan solo con el grado terceroo.
- Estrato 2: escuelas que cuentan con el grado terceroo y sexto conjuntamente.
- Estrato 3: escuelas que cuentan solo con el grado sexto.

Ahora bien, nótese que para seleccionar las escuelas dentro de cada estrato, se utiliza un diseño proporcional, cuya medida de tamaño es el número de estudiantes en los grados terceroo y sexto. Teniendo esto en mente, para construir un diseño eficiente y que sea representativo, debemos garantizar que

$$\sum_s w_k^{(design)} = \hat{N} \approx N$$

.

Por ejemplo, para el caso del grado terceroo, si asumimos que $N_3^{schools}$ representa el total de estudiantes de este grado, entonces:

$$\sum_{i \in s_{36}} w_{2i}^{(school)} + \sum_{i \in s_3} w_{3i}^{(school)} \approx N_3^{schools}$$

En donde $w_{2i}^{(school)} = \frac{t_{x2}}{n_{I2} x_i}$ es el peso de la escuela en el Estrato 2; $w_{3i}^{(school)} = \frac{t_{x3}}{n_{I3} x_i}$ es el peso de la escuela en el Estrato 3; t_{x2} es el total de estudiantes de grado terceroo y sexto; t_{x3} total de estudiantes del grado sexto; n_{Ii} es el tamaño de la muestra seleccionada en el estrato i con $i = 1, 2, 3$; s_{36} es la muestra obtenida en el Estrato 2 y s_3 es la muestra obtenida en el Estrato 1.

6.3 Ajuste por ausencia de respuesta

Las tasas de respuesta de escuelas y estudiantes no siempre corresponden al 100% de la población muestreada. La ausencia de respuesta en ambas unidades afecta los análisis estadísticos, razón por la cual deben ajustarse los pesos de muestreo por medio de factores que disminuyan el sesgo ocasionado por la ausencia de respuesta. Estos factores deben ser una función de una variable que

esté disponible para aquellos que contestaron y para los que no, y que esté correlacionada con la tasa de respuesta y las variables de interés.

Para encontrar esta variable es posible buscar variables que estén relacionadas con el hecho de responder o no a la prueba y esperar que estén relacionadas también con las variables de interés (teniendo en cuenta los datos del estudio), o buscar variables que cumplan ambos requisitos con seguridad (considerando datos históricos de otro estudio que tenga características similares).

Esta última técnica es la más utilizada, puesto que ofrece más estabilidad (las tasas de respuesta pueden variar mucho en cada aplicación), permite la estandarización de procesos, lo cual es útil para generar ajustes en cortos periodos de tiempo, y facilita la identificación y uso de las variables de mayor interés en el estudio, como el logro de aprendizaje que es evaluado. Una vez definida la variable los ajustes se aplican a las escuelas y, posteriormente, a los estudiantes.

Generalmente en una investigación por muestreo se presentan dos tipos de ausencia de respuesta: una es cuando la persona seleccionada se niega a responder algunas preguntas del cuestionario, la otra es cuando hay imposibilidad de levantar toda la información del cuestionario por cualquier razón. El primer tipo de ausencia de respuesta se puede solucionar a través de métodos de imputación que buscan asignar información plausible a las variables no levantadas en campo, usando la información de individuos similares. La mayoría de textos de análisis de datos presentan los métodos de imputación con sus ventajas y desventajas; lo más importante es controlar adecuadamente los niveles de ausencia de respuesta, pues se aconseja que los datos imputados no tengan una tasa alta.

Särndal y Lundstrom (2005) muestran diferentes métodos para el tratamiento de la ausencia de respuesta, donde además se enfatiza en los supuestos para poder llevar a cabo este tipo de procedimientos de imputación de datos entre los cuales se destaca que el esquema de datos faltantes sea aleatorio. El tratamiento de la ausencia de respuesta implicará el uso de estimadores de calibración, como información auxiliar se considerará el total de estudiantes y demás parámetros disponibles para este grupo poblacional, si es posible.

6.3.1 Ausencia de respuesta de escuelas

La ausencia de respuesta en las escuelas no está relacionada fuertemente con los logros de aprendizaje, puesto que estas tasas de respuesta se ven afectadas por otros temas del contexto de cada país que no afectan el rendimiento promedio de los estudiantes. Sin embargo, las tasas de respuesta pueden variar mucho entre los estratos y países, por lo cual un ajuste es necesario para mitigar el sesgo. Además, si una escuela originalmente muestreada decide no participar y sus dos reemplazos no acceden a ser parte del estudio, el peso de la escuela debe ajustarse para compensar la pérdida. El factor de ajuste por ausencia de respuesta permite ponderar las escuelas para representar a todos los estudiantes. El mecanismo es crear distintos grupos de escuelas con características similares y, dentro de cada grupo, ajustar los pesos de las escuelas para compensar aquellas sin respuesta.

Las variables utilizadas para realizar el ajuste son, por lo general, variables de estratificación, porque están relacionadas con las variables de interés y representan subpoblaciones de interés para

los países participantes. De igual forma, en el cálculo del factor de ajuste se considera la matrícula estimada de estudiantes para cada escuela, con el fin de reconocer el número de estudiantes que están representados por cada escuela en la muestra.

6.3.2 Ausencia de respuesta de estudiantes

La ausencia de respuesta de los estudiantes es, por lo general, mayor que la observada en las escuelas y, de igual forma, se relaciona en mayor medida con los logros de aprendizaje. Con frecuencia, los estudiantes que se ausentan de la prueba son aquellos con el rendimiento académico más bajo, quienes no muestran interés en asistir a la escuela o que tienen alguna condición de salud deficiente, lo cual puede causar un sesgo en las estimaciones, presentando resultados más altos que los reales. El ajuste por ausencia de respuesta de estudiantes disminuye ese sesgo.

Debido a que la información sobre las características de los estudiantes es usualmente limitada, es posible utilizar otras variables (de la escuela) como predictores de la respuesta del estudiante. Así, los ajustes se realizan formando grupos de estudiantes de escuelas similares: se puede comparar estudiantes en escuelas públicas y privadas o de alto y bajo desempeño, por ejemplo.

6.3.3 Ausencia de respuesta en cuestionarios de contexto

Es posible que los cuestionarios de contexto (no cognitivos) - que indagan al estudiante, profesor o rector acerca de los factores asociados al desempeño escolar - no sean diligenciados en su totalidad por los respondientes. En este caso, es posible considerar procedimientos de imputación. Son varios los autores que han propuesto diferentes procedimientos para tratar la ausencia de respuesta, existen diversos paquetes en `R` que permiten modelar este fenómeno. Por ejemplo, los paquetes `MissMDA`, `mice`, `statmatch`, `hotdeck`, ‘mitools entre otros. Sin embargo, el propósito debe ser siempre garantizar que el método que se utiliza es apropiado y se ajusta bien al esquema de ausencia de respuesta del estudio. Por esta razón se propone realizar una prueba usando una muestra de entrenamiento con la cual se genera el modelo de imputación, se deja una muestra de comprobación donde no se tenga valores faltantes, con esta se calcula la tasa de error aparente con el fin de identificar si el modelo de imputación es apropiado o haciendo pruebas de hipótesis sobre la distribución de los datos sin imputar e imputados.

Para el tratamiento de la ausencia de respuesta, como se mencionó, se puede usar la función mice del paquete de R con el mismo nombre, considerando el método que viene por defecto, el método de predicción de emparejamiento por la media (Predictive mean matching). Los detalles de este método de imputación se describen en van Buuren (2012).

En el segundo tipo de ausencia de respuesta, se da cuando la sobremuestra generada no logra cubrir la ausencia de respuesta y esta se considera no ignorable por las características de las unidades finales de muestreo que no respondieron, se debe llevar a cabo un ajuste a los factores de expansión.

El ajuste de los factores de expansión busca fundamentalmente evitar sesgos debidos a la ausencia de respuesta.

6.4 Estimadores ajustados

La falta de respuesta de unidades finales de muestreo no ignorables obliga a realizar una calibración o ajuste al factor inicial calculado en el diseño, a través del cálculo de un factor de ajuste. El peso muestral total de los estudiantes se define como el producto de los pesos de muestreo de las escuelas y de los estudiantes, después de ajustar por ausencia de respuesta y recortes por estrato. Este resultado permite calcular de forma precisa los estimadores de las características de la población objetivo. La expresión matemática para el cálculo del factor de expansión final es:

$$w_k^{(nonr)} = w_k^{(design)} * w_k^{(adj)}$$

En donde $w_k^{(design)}$ hace referencia al peso de muestreo inducido por el diseño de muestreo aplicado en cada país y $w_k^{(adj)}$ hace referencia al factor de ajuste que debe realizarse para modelar la ausencia de respuesta (no ignorable) en cada país.

El principal reto para la calibración del factor debido a la ausencia de respuesta es que implica disponer de información secundaria para las unidades finales de muestreo, de modo que los resultados de los que efectivamente respondieron sean extrapolables al universo; aún en el caso en que haya una fracción de éstos que inicialmente fueron seleccionados en la muestra y que no respondieron o no pudieron ser contactados. En esté documento se propone utilizar la metodología definida por Gutiérrez & Rojas (2013) donde se ajusta la forma funcional del estimador final utilizando la metodología de *propensity score*.

Para esto, se denota la muestra de los respondientes como s_r y para cada individuo $k \in s$ se observa la variable R_k, la cual es una variable indicadora para el evento *el k-ésimo individuo es respondiente*. Por lo anterior, R_k tiene distribución Bernoulli con parámetro

$$\phi_k = Pr(R_k = 1) = Pr(k \in s_r)$$

Por otro lado, para estimar correctamente ϕ_k, se debe contar con un vector de información auxiliar $\mathbf{z}_k$ conocido para todo $k \in s$ se puede estimar por medio de un modelo de regresión logística; esto es,

$$\hat{\phi}_k = \frac{\exp\{\mathbf{z}_k'\hat{\beta}\}}{1 + \exp\{\mathbf{z}_k'\hat{\beta}\}}$$

donde $\hat{\beta}$ es el vector de coeficientes estimado de la regresión logística. Finalmente, el estimador para un total poblacional, con el ajuste debido a la ausencia de respuesta no ignorable, queda expresado como

$$\hat{t}_{adj} = \sum_{k \in s_r} w_k^{(nonr)} y_k$$

En donde

$$w_k^{(nonr)} = w_k^{(design)} * w_k^{(adj)} = \frac{w_k^{(design)}}{\hat{\phi}_k}$$

La varianza de este estimador está dada por

$$\widehat{Var}(\hat{t}_{adj}) = \frac{1}{N^2} \sum_{k \in s_r} \frac{(1 - \hat{\phi}_k)}{\pi_k \hat{\phi}_k^2} (y_k - \mathbf{z}_k' \hat{\phi}_k \hat{\gamma})^2$$

En donde la expresión para $\hat{\gamma}$ puede ser consultada en Kim & Riddles (2012). En ausencia de información auxiliar que permita establecer una buena estrategia de modelamiento para la ausencia de respuesta, es posible recurrir a métodos de ajuste tradicionales como por ejemplo el que Bautista (1998) presenta, en donde se supone un factor de ajuste clásico para corregir defectos del marco, cobertura y de ausencia de respuesta, el cual tiene la siguiente expresión matemática:

$$w_k^{(adj)} = \frac{n + n_{add} - n_{out}}{n + n_{add} - n_{out} - n_{rej}}$$

donde n es el tamaño de muestra propuesto, n_{add} es la cantidad de elementos adicionales que se encuentran por defectos del marco, n_{out} es la cantidad de elementos que están fuera del universo de estudiantes y n_{rej} es la cantidad de rechazos en la muestra.

6.5 Calibración de los pesos de muestreo iniciales

Después de considerar el ajuste por ausencia de respuesta, es posible calibrar los pesos de muestreo sobre la información auxiliar disponible en los sistemas educativos de cada país, a nivel nacional, por estratos de interés, e incluso por las sobremuestras definidas por los coordinadores nacionales.

Es posible utilizar un enfoque de calibración funcional para el ajuste de los factores de expansión bajo información faltante. Así, Särndal y Lundström (2005) afirman que cuando los estudios por muestreo están afectados por la ausencia de respuesta, es deseable tener las siguientes propiedades en la estructura inferencial que sustenta el muestreo:

1. Sesgo pequeño o nulo.

2. Errores estándares pequeños.
3. Un sistema de ponderación que reproduzca la información auxiliar disponible[2].
4. Un sistema de ponderación que sea eficiente al momento de estimar cualquier característica de interés en un estudio multipropósito.

Las anteriores características son satisfechas al usar el enfoque de calibración (Deville, Särndal, Swensson & Wretman, 2003) que induce una estructura inferencial robusta en presencia de información auxiliar precisa puesto que reduce tanto el error de muestreo como el error debido a la ausencia de respuesta. Por lo anterior, se considera que existe información auxiliar disponible para cada respondiente. Como vector de información auxiliar $\mathbf{t_Z} = (t_z^*, \hat{t}_z^0)'$, en donde $t_z^* = \sum_{k\in U} z_k^*$ se define como el total auxiliar disponible a nivel de la población finita U, y $\hat{t}_z^0 = \sum_{k\in s} \frac{z_k^0}{\pi_k}$ representa la información auxiliar disponible a nivel de la muestra s, expandida a la población mediante el estimador de Horvitz-Thompson.

Esta propuesta considera un procedimiento en dos etapas, en donde primero se calibra el conjunto de pesos ajustados, a partir de la información auxiliar z_k^0. Esta primera etapa calibra desde la muestra efectiva de respondientes s_r a la muestra original s. Luego, se realiza una segunda calibración, a partir de la información auxiliar z_k^*, desde la muestra s a la población de interés U. Luego, como pesos finales se considera el siguiente sistema

$$w_k^{(cal)} = w_k^0 v_k^*$$

En donde,

$$v_k^* = 1 + \lambda_r' \mathbf{z}_k^*$$

representa el factor de ajuste para la calibración. Nótese que $\lambda_r' = (\sum_U \mathbf{z}_k^* - \sum_{s_r} w_k^0 \mathbf{z}_k^*)'(\sum s_r w_k^0 \mathbf{z}_k^* \mathbf{z'}^*_k)^{-1}$.

Además,

$$w_k^0 = w_k^{(nonr)} v_k^0$$

En donde, $v_k^0 = 1 + \lambda_r'^0 \mathbf{z}_k^0$ y $\lambda_r'^0 = (\sum_s w_k^{(nonr)} \mathbf{z}_k^0 - \sum_{s_r} w_k^{(nonr)} \mathbf{z}_k^0)'(\sum s_r w_k^{(nonr)} \mathbf{z}_k^0 \mathbf{z'}^0_k)^{-1}$. Luego, la estimación del total calibrado se calcula como:

$$\hat{t}_{cal} = \sum_{k\in s_r} w_k^{(cal)} y_k$$

Para implementar los estimadores de calibración es posible utilizar las siguientes funciones implementadas en el software estadístico R:

- `calibrate` del paquete `survey` (Lumley, 2017).

[2] Por ejemplo, el número de escuelas en el país, o el número de estudiantes en el sistema educativo.

- `Wk` del paquete `TeachingSampling` (Gutiérrez, 2017).
- `calib` del paquete `sampling` (Tillé & Matei, 2016).

Capítulo 7

Cálculo de la varianza muestral

Las fórmulas computacionales requeridas para estimar la varianza de estadísticas descriptivas como la media muestral están disponibles para algunos diseños complejos que incorporan elementos como la estratificación y el muestreo por conglomerados. Sin embargo, en el caso de estadísticas analíticas más complejas, tales como coeficientes de correlación y coeficientes de regresión, no se encuentra fácilmente las fórmulas específicas en diseños muestrales que se aparten del muestreo aleatorio simple. Estas fórmulas son enormemente complicadas o, en última instancia, se resisten al análisis matemático (Frankel, 1971).

En ausencia de fórmulas adecuadas, en los últimos años han aparecido una variedad de técnicas empíricas que proporcionan *varianzas aproximadas que parecen satisfactorias para fines prácticos* (Kish, 1995). Estos métodos utilizan una muestra de datos para construir submuestras y generar una distribución para las estimaciones de los parámetros de interés utilizando cada submuestra. Los resultados de la submuestra se analizan para obtener una estimación del parámetro, así como intervalos de confianza para esa estimación.

De forma alternativa, la aproximación en series de Taylor se puede utilizar para proporcionar un método de estimación de la varianza más “directo” que los proporcionados por los tres enfoques mencionados previamente. Las series de Taylor se utilizan para aproximar numéricamente los primeros términos de una expansión en serie de la fórmula de varianza. En la actualidad existe un buen número de programas informáticos que fueron diseñados para llevar a cabo los intensivos cálculos numéricos requeridos por el enfoque (Wolter, 1985).

7.1 La técnica de Jackknife

Es posible estimar la varianza de los estimadores de interés usando la técnica de Jackknife. El desarrollo del procedimiento de Jackknife se remonta a un método utilizado por Quenouille (1956) para reducir el sesgo de las estimaciones. El refinamiento ulterior del método (Mosteller & Tukey, 1968)

llevó a su aplicación en una serie de situaciones de las ciencias sociales en las que las fórmulas no están fácilmente disponibles para el cálculo de errores de muestreo.

Este procedimiento ofrece los siguientes beneficios:

1. *Mayor flexibilidad*: el Jackknife puede implementarse en una amplia variedad de diseños muestrales.
2. *Facilidad de uso*: el Jackknife no requiere de software especializado.

El concepto principal de esta técnica parte de una muestra de tamaño n, la cual se divide en A grupos de igual tamaño $m = n/A$, a partir de esta división, la varianza de un estimador $\hat{\theta}$ se estima a partir de la varianza observada en los A grupos.

Para cada grupo $(a = 1, 2, ..., A)$, se calcula $\hat{\theta}_{(a)}$, una estimación para el parámetro θ, calculada de la misma forma que la estimación $\hat{\theta}$ obtenida con la muestra completa, pero solo con la información restante (luego de la eliminación del grupo a). Para $a = 1, 2, ..., A$ se define

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_{(a)}$$

como un pseudovalor de θ. El estimador obtenido mediante Jackknife se presenta como una alternativa a $\hat{\theta}$ y se define como:

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a$$

mientras que el estimador de la varianza obtenido mediante Jackknife se obtiene como:

$$\hat{V}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2$$

También es posible utilizar como estimador alternativo:

$$\hat{V}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2$$

Para diseños estratificados y multietápicos en los cuales unidades primarias de muestreo han sido seleccionadas en el estrato h, para $h = 1, \ldots, H$, el estimador de varianza de Jackknife para la estimación de un parámetro poblacional está dado por

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H}\frac{n_h - 1}{n_h}\sum_{i=1}^{n_h}(\hat{\theta}_{(hi)} - \hat{\theta})^2$$

donde $\hat{\theta}_{(hi)}$ es la estimación de θ usando los datos de la muestra excluyendo las observaciones en la i-ésima unidad primaria de muestreo (Korn & Graubard, 1999, pg. 29 – 30). Shao & Tu (1995, Teorema 6.2) garantiza la convergencia en probabilidad de este estimador hacia la varianza teórica, de donde se puede concluir que es un estimador aproximadamente insesgado para la varianza teórica. Computacionalmente, se puede obtener la estimación Jackknife por medio de la creación de la base de datos omitiendo las observaciones necesarias usando comando `as.svrepdesign` de la librería `survey` del software estadístico 'R para calcular $\hat{V}_{JK}(\hat{\theta})$, y posteriormente calcular el valor de la estimación Jackknife.

7.2 El método de las Réplicas Repetidas Balanceadas

Las varianzas de muestreo pueden ser calculadas haciendo uso del método conocido como Réplicas Repetidas Balanceadas, el cual permite explicar la varianza que se obtiene en las estimaciones debido al muestreo. Este método es el que utilizan pruebas internacionales como PISA para realizar los análisis de datos.

Para la aplicación de la Réplicas Repetidas Balanceadas es recomendable usar el método de Fay, el cual es similar al método Jackknife, pero es más apropiado cuando hay funciones no diferenciables en el estudio. En PISA, por ejemplo, el método de Fay es preferido porque el método Jackknife no proporciona un estimador de varianza estadísticamente consistente para los cuantiles. Por otro lado, la Réplicas Repetidas Balanceadas brinda estimadores lineales simples que son imparciales y consistentes. Además, tiene una consistencia asintótica deseable para un conjunto amplio de estimadores, bajo diseños complejos y estudios de simulación empírica.

Para la aplicación de este método, los pesos de muestreo se ajustan para generar los pesos de repetición y, posteriormente, se repiten los ajustes por ausencia de respuesta de escuelas y estudiantes para estos nuevos pesos. Con estos pesos de repetición se estiman los errores de muestreo y la varianza de muestreo, incluyendo el impacto de la ausencia de respuesta, el cual se espera que sea pequeño, pero relevante en el momento de calcular estimadores más precisos.

En TXXXXXXX se aplicó este método de la siguiente manera: primero, las unidades de muestreo (escuelas y estudiantes) fueron agrupadas en pares y se construyeron unos pseudo-estratos; segundo, dentro de cada pseudo-estrato se eliminó una de las unidades (siguiendo una matriz de Hadamard) y se recalculó el peso (peso replicado) para la otra; y terceroo, para cada conjunto de pesos replicados se obtuvo el estadístico de interés y se determinó su error estándar.

7.2.1 Modificación de Fay

La técnica de Fay sigue el mismo proceso que la técnica de Jackknife para identificar las parejas de escuelas dentro de cada estrato; sin embargo, no elimina una de las dos escuelas, sino que pondera los pesos muestrales de ambas utilizando un coeficiente entre 0 y 1 (la suma de los coeficientes de las escuelas debe ser 1), lo cual permite obtener estimaciones más robustas para los errores estándar.

En TXXXXXXX los coeficientes que se usaron fueron 1,5 para una escuela y 0,5 para la otra. Las entradas de una matriz de Hadamard[1] de tamaño 100 y que toma valores de +/- 1 fueron utilizadas para definir el coeficiente asignado a cada escuela. Para este estudio se utilizaron dos procesos diferentes en el caso de escuelas y estudiantes: el peso de muestreo ajustado por ausencia de respuesta para las escuelas se multiplicó por el coeficiente establecido para obtener los ponderadores finales; mientras que, en el caso de los estudiantes, se utilizó el ponderador final de las escuelas para generar ponderadores de replicación, que fueron ajustados por la ausencia de respuesta de estudiantes después de haber sido definidos.

7.3 La estimación de la varianza con valores plausibles

Debido a que los estudiantes responden solo un subconjunto de ítems de la totalidad de ítems desarrollados para la prueba, la estimación de su habilidad individual está sujeta a un error de medición. Los enfoques tradicionales de estimación de la habilidad, como la máxima verosimilitud marginal (MML) y estadística bayesiana, arrojan estimaciones óptimas para la habilidad individual, pero sesgadas a nivel de grupo. Una forma de tener en cuenta la incertidumbre asociada a las estimaciones y de obtener estimaciones insesgadas a nivel de grupo, es a partir del uso de múltiples valores plausibles que representan la distribución probable de la habilidad de un estudiante.

El desempeño de los estudiantes que presentan las pruebas estandarizadas, es obtenido a partir de cinco valores (denominados valores plausibles). Estos valores son útiles, pues tienen en cuenta la aleatoriedad producida por el hecho de que los estudiantes responden a un número pequeño de preguntas (lo cual permite obtener mejores estimaciones de las estadísticas de interés relacionadas con el desempeño en las pruebas a nivel agregado) y no todos los estudiantes responde el mismo conjunto de ítems (von Davier, M., González, E. & Mislevy, R., 2009).

Por ejemplo, es posible modelar la probabilidad de que un indiviudo j de habilidad θ_j conteste acertadamente el item i. En particular, al asumir un modelo de califiación logístico de tres parámetros (3PL), el modelo toma la siguiente forma:

$$P(U_i = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp\{a_i(\theta_j - b_i)\}}{1 + \exp\{a_i(\theta_j - b_i)\}}$$

En donde para el item i, se tiene que a_i representa el parámetro de discriminación; b_i, el parámetro de dificultad y c_i, el parámetro asociado al pseudo-azar. Ahora, los valores plausibles se obtienen como una muestra aleatoria de la distribución condicional de la habilidad para cada estudiante, denotada por la siguiente expresión

[1] La matriz de Hadamard se caracteriza porque el producto vectorial de una fila y una columna del mismo número es igual al rango de la matriz y el producto de cualquier fila con una columna de un número distinto es cero. Esta matriz se construye fácilmente, teniendo en cuenta que sus dimensiones son un múltiplo de cuatro (ver Rutkowski, L., von Davier, M. & Rutkowski, D. (2014), pg. 147).

$$P(\theta_j|U_i, a_i, b_i, c_i, \mathbf{Z}, \beta, \sigma^2)$$

Nótese que en esta se condiciona tanto por las respuestas dadas por el evaluado a los ítems y los parámetros estimados para el modelo de Teoría de Respuesta al Ítem 3PL (a_i, b_i, c_i), como por toda la información auxiliar disponible para cada evaluado ($\mathbf{Z}$), y sus relaciones con la habilidad de los evaluados en la población (β) y una varianza común a todos los evaluados (σ^2).

Dada la estimación de los parámetros para el modelo, es necesario conocer los valores de β y σ^2 para definir la distribución condicional de la habilidad. Para esto, se asume un modelo lineal para la habilidad condicional $P(\theta_j|\mathbf{Z}_j, \beta, \sigma^2) = N(\mathbf{Z}_j, \beta, \sigma^2)$ la cual puede ser expresada como:

$$P(\theta_j|U_i, a_i, b_i, c_i, \mathbf{Z}_j, \beta, \sigma^2) \approx P(U_i|\theta_j, a_i, b_i, c_i)P(\theta_j|\mathbf{Z}_j, \beta, \sigma^2)$$

en la cual, los parámetros se estiman mediante un algoritmo de Esperanza-Maximización. Para realizar estimaciones relacionadas con los puntajes de los estudiantes con base en los valores plausibles, se debe realizar la estimación de la estadística de manera independiente para cada uno de los valores plausibles, teniendo en cuenta, además, los factores de expansión correspondientes. De manera que, suponiendo que la estadística que se desea estimar es η, se deben encontrar las estimaciones $\hat{\eta}_1, \cdots, \hat{\eta}_5$ para los cinco valores plausibles, y la estimación de interés se obtiene de la forma:

$$\hat{\eta} = \frac{1}{5}\sum_{i=1}^{5} \hat{\eta}_i$$

Por ejemplo, si se desea estimar el puntaje promedio de todos los estudiantes, se estima el promedio con cada uno de los 5 valores plausibles. Al final se obtienen cinco estimaciones, las cuales se promedian para determinar la estimación del puntaje promedio. Los errores estándar asociados a las estimaciones que emplean valores plausibles se obtienen a partir de la raíz cuadrada de la varianza de las estimaciones, para la cual se toman en cuenta los siguientes dos componentes:

1. La varianza de la estimación debida al diseño muestral.
2. La varianza debida a la incertidumbre de la medición.

Así, contemplando el uso de cinco valores plausibles, el error estándar asociado con una estimación se calcula como:

$$EE(\hat{\eta}) = \sqrt{\frac{1}{5}\sum_{i=1}^{5} Var_m(\hat{\eta}_i) + \left(1 + \frac{1}{5}\right)\frac{1}{4}\sum_{i=1}^{5}(\hat{\eta}_i - \hat{\eta})^2}$$

Donde $Var_m(\hat{\eta}_i)$ es la varianza del estimador calculada con el i-ésimo conjunto de valores plausibles y de acuerdo con el diseño muestral.

El número total de réplicas L consideradas para la estimación de la varianza del estimador es igual al número de establecimientos en la muestra. Si se denomina $\hat{\eta}_j$ a cada una de las estimaciones con

las réplicas y $\hat{\eta}$ a la estimación con la muestra completa, la estimación de la varianza del estimador queda expresada como

$$Var_m(\eta) = \sum_{l=1}^{L} \frac{n_l - 1}{n_l} \sum_{j=1}^{n_l} (\hat{\eta}_j - \hat{\eta})^2$$

Este tipo de estimadores de varianza dan una gran flexibilidad al momento de estimar diferentes tipos de parámetros (promedios, pXXXXXXXntiles, porcentajes), ya que el esquema planteado es transversal ante cualquier diseño de muestreo.

7.4 Detalles computacionales

El software que se usará para obtener las estimaciones de varianza es ℝ; el paquete `survey` (Lumley, 2017) contiene funciones que permiten estimar la varianza de estimadores como totales, proporciones, razones y promedios de diseños muestrales complejos.

La función `svydesign` permite ingresar la información de diseño de muestreo elegido (todas las etapas de muestreo y los aspectos de estratificación). Esta función tiene programadas las fórmulas matemáticas para lograr una aproximación de la estimación de la varianza a partir de la linealización de Taylor.

Capítulo 8

Referencias

- Bautista, L. (1998). Diseños de Muestreo Estadístico. Universidad Nacional de Colombia.
- Caro, D. & Biecek, P. (2015),intsvy: An R Package for Analysing International Large-Scale Assessment Data.
- Cervantes, V. H., Cepeda, E. & Camargo, S. L. (2008), 'Una propuesta para la obtención de niveles de desempeño en los modelos de teoría de respuesta al ítem', Avances en Medición 6(1), 49-58.
- Elosua, P. (2011), Introducción al entorno R. Argitalpen Zerbitzua.
- Fernández, J. M. (1997), Introducción a la teoría de respuesta a los ítems, Ediciones Pirámide.
- Frankel (1971). Inference from Survey Samples. Michigan University.
- Greaney, V. & Kellaghan, T. (2016), Volumen 3, Implementación de una evaluación nacional del rendimientoacadémico. Grupo Banco Mundial.
- Gutiérrez, A. (2015), Estrategias de muestreo, Diseño de encuestas y estimación de parámetros. Universidad Santo Tomás.
- Gutiérrez, A. (2017), TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population. R package version 3.2.2. https://CRAN.R-project.org/package=TeachingSampling
- Gutiérrez, A. (2017), samplesize4surveys: Sample Size Calculations for Complex Surveys. R package version 3.1.2.400. https://CRAN.R-project.org/package=samplesize4surveys
- Gutiérrez, A. and Rojas, L. (2013), Ajuste de estimadores mediante la técnica de propensity score en encuestas complejas. Revista IB (DANE).
- Joncas, M. & Foy, P. (2013), Sample Design in TIMSS and PIRLS.
- Kim, J. and Riddles, F. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. Survey Methodology.

- Kish, L (1995). Survey Sampling. Wiley.
- Korn, E., and Graubard, B. (1999) Analysis of Health Surveys. Wiley.
- Lazarsfeld, P. F. (1955), Recent developments in latent structure analysis. Sociometry 18(4), 391-403.
- Lord, F. M. & Novick, M. R. (1968), Statistical theories of mental test scores, Information Age Publishing.
- Lumley, T. (2004). Analysis of complex survey samples. Wiley.
- Lumley, T. (2017). survey: analysis of complex survey samples. R package version 3.32.
- Martin, M., Mullis, I. & Hooper, M. (2016), Methods and Procedures in TIMSS 2015. International Association for the Evaluation of Educational Achievement (IEA).
- Mosteller, F., and Tukey, J. (1968). Data analysis, including statistics. Handbook of Social Psychology.
- PISA (2009), PISA Data Analysis Manual. SPSS, Second Edition. OECD Publishing.
- PISA (2012), Technical Report. OECD Publishing.
- PISA (2016), Annex A6, The PISA 2015 Field Trial Mode-Effect Study. OECD Publishing.
- PISA (2017), Main Survey School Sampling Preparation Manual, Overview. OECD Publishing.
- PISA-based Test for Schools (2017), Technical Report (Draft version). OECD Publishing.
- Quenouille, M. H. (1956), Notes on bias in estimation, Biometrika.
- Ross, K. (2005), Module 1, Educational research: some basic concepts and terminology. UNESCO.
- Ross, K. (2005), Module 3, Sample design for educational survey research. UNESCO.
- Rutkowski, L., von Davier, M. & Rutkowski, D. (2014), Handbook of International Large-Scale Assessment. Background, Technical Issues, and Methods of Data Analysis. CRC.
- Särndal, Swensson & Wretman (2003). Model Assisted Survey Sampling. Springer.
- Särndal y Lundstrom (2005). Estimation in Surveys with Nonresponse. Wiley.
- Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Springer-Verlag
- TERCE (2015), Informe de Resultados Tercer Estudio Regional Comparativo y Explicativo. Logros de Aprendizaje. Ediciones UNESCO.
- TERCE (2015), Informe de Resultados Tercer Estudio Regional Comparativo y Explicativo. Factores Asociados. Ediciones UNESCO.
- Tillé, Y. and Matei, A. (2016). sampling: Survey Sampling. R package version 2.8. https://CRAN.R-project.org/package=sampling

- Tukey, J. (1949), ‘Bias and confidence in not quite-large samples’, Annals of Mathematical Statistics 2(29), 614.
- UNESCO – OREALC (2016), Reporte Técnico TERCE. Ediciones UNESCO.
- van Buuren (2012). Flexible Imputation of Missing Data. CRC Press.
- von Davier, M., and Gonzalez, E., and Mislevy, R. (2009). What are Plausible Values and Why are They Useful?. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments.
- Westat (2007), WesVar 4.3, user’s guide.
- Wolter, K. (1985). Introduction to Variance Estimation. Springer.

Bibliografía