NOTA TÉCNICA

Definición de niveles de desarrollo para las escalas del instrumento de medición de habilidades socioemocionales en el contexto escolar colombiano usando la Teoría de Respuesta al Ítem[1]

31 de agosto de 2020

[1] Autores: Jorge Luis Bazan, profesor asociado del Departamento de Matemáticas Aplicadas de la Universidad de Sao Paulo (jorgeluisbazan@gmail.com); Juan Camilo Suárez, economista consultor en educación (juan.suarezgo@gmail.com) y Lina Ramírez, economista consultora en educación (lmramirezvillegas@gmail.com).

Tabla de contenido

Introducción

El presente estudio se entiende como una continuación del estudio de "*Validación de un instrumento para medir habilidades socioemocionales en el contexto escolar colombiano, 2020".* Cuyo propósito consistió en diseñar, adaptar, pilotear y validar psicométricamente un inventario de escalas que permitiera realizar medición del desarrollo de habilidades socioemocionales en niños, niñas y jóvenes en edad escolar.

Los resultados del primer estudio fueron satisfactorios acerca de las caracteristicas psicometricas del instrumento y dieron lugar a un instrumento que cuenta con tres versiones: i) inventario para educación básica primaria conformado por 15 escalas; ii) inventario para educación básica secundaria conformado por 16 escalas, e; iii) inventario para educación media, conformada por 18 escalas. El instrumento cubre un amplio espectro de habilidades que se enmarcan en la perspectiva conceptual propuesta por la Colaboración para el aprendizaje académico, social y emocional (CASEL por sus siglas inglés). Las escalas abordan tres grandes temáticas a saber: habilidades emocionales (permiten entender y manejar nuestras emociones), habilidades sociales (permiten construir y mantener relaciones positivas y duraderas con otras personas) y la toma de decisiones de manera responsable (capacidad de tomar decisiones responsables consigo mismo y con los demás). Para cada versión, las escalas son confiables y unidimensionales.

En cuanto a la muestra del estudio piloto se resalta el hecho de que no fue aleatoria ni representativa de la población escolar de Colombia y se construyó principalmente en colaboración de las instituciones educativas con las cuáles la Fundación Corona, la Fundación Luker, la Alianza Educativa[2] y la Fundación Carvajal adelantan intervenciones en educación[3]. La muestra final utilizada para evaluar las propiedades psicométricas del instrumento fue obtenida de cerca de seis mil estudiantes (aproximadamente dos mil estudiantes por cada versión del instrumento) que asistían, en el segundo semestre de 2019, a 41 sedes y 31 instituciones educativas ubicados en 6 departamentos y 10 municipios del país.

A pesar de no ser representativa, la muestra fue suficiente para evaluar satisfactoriamente la confiablidad y la unidimensionalidad de la totalidad de las escalas de las tres versiones del instrumento. En general, todas las escalas resultaron tener buenas propiedades psicométricas a excepción de autorregulación y postergación de la gratificación (en todos los niveles educativos) y autopercepción general (en básica primaria). Los resultados también demostraron ser homogéneos entre distintos subgrupos de la población (genero, grado y origen étnico)

El primer estudio también tenía como objetivo proponer una metodología para calcular puntajes y establecer puntos de corte que permitiera a los usuarios del instrumento clasificar a

[2] La Alianza Educativa participó en el piloto a través de varios de los colegios que gestiona a través del modelo de administración del servicio educativo creado por la Secretaría de Educación Distrital de Bogotá.

[3] Otras de las organizaciones que facilitaron el contacto con colegios son: La Secretaría de Educación de Bogotá y el Instituto Colombiano para la Evaluación de la Educación (ICFES).

los estudiantes en niveles de desarrollo en función de los puntajes obtenidos en cada escala (ver Protocolo de implementación, 2020). Debido a la naturaleza ordinal del formato de respuesta de las escalas, el referencial teórico elegido para el análisis fue el de Teoría Clásica de los Tests (TCT) que consiste en sumar los puntajes de las preguntas para obtener un puntaje total de cada escala. Por otro lado, la definición de puntos de corte se realizó con la asesoría de un psicólogo experto y a través de un enfoque basado en criterios que se centra en la interpretación del constructo.

Aunque el enfoque escogido para el cálculo de medidas y la definición de puntos de corte es legítimo y ha sido ampliamente utilizado en la literatura psicométrica, su uso puede tener varias limitaciones, entre ellas el hecho que las preguntas se asumen igualmente importantes (tienen el mismo peso) y al ser ordinales en sus respuestas se asumen como de tipo de intervalo al sumarlas (Bazán, 2014). En la actualidad existen enfoques que han demostrado poseer mejores propiedades en la medición psicométrica, es el caso de la Teoría de Respuesta al Ítem (TRI). Dentro de las ventajas del uso de TRI frente a TCT está el hecho de que los parámetros de las preguntas son estables aunque las personas que contesten sean distintas y que el nivel de medida de interés obtenido para cada individuo está basado en el tipo de respuestas que esté da y no necesariamente al número de respuestas.

Por tanto, en vista de las ventajas que el uso de TRI tiene para mejorar la precisión, interpretabilidad y comparabilidad de resultados de futuras aplicaciones del instrumento, este estudio tiene como objetivo usar este enfoque para calcular las medidas y estimar los puntos de corte de las escalas. Es importante resaltar que el uso de este método psicométrico moderno requiere del cumplimiento de algunos supuestos asociados a las medidas. Así las cosas, su aplicación se hace posible debido a los buenos resultados obtenidos en el primer estudio, en dónde se hallaron evidencias estadísticas en favor de la unidimensionalidad de las escalas.

El instrumento está conformado por preguntas que tienen dos formatos de categorías de respuesta: escalas dicotomicas o binarias y escalas politómicas (con más de dos altenativas de respuesta de tipo ordinal). Para ambos casos hay diferentes modelos dentro del contexto de la Teoría de Respuesta al Ítem. Específicamente el Modelo de Respuesta Graduada (MGR) propuesto por Samejima (1969) resulta ser el modelo más apropiado para las características de las escalas del instrumento[4]. Debido al reciente desarrollo computacional, el MRG es cada vez más usado para las escalas ordinales (ver por ejemplo Nering y Ostini, 2011). En este sentido, el estudio adoptará este modelo TRI para la estimación de puntajes y puntos de corte.

La aplicación de MRG consta de dos etapas secuenciales pero independientes, usualmente denominadas de *calibración* de preguntas y de *estimación* de las medidas. La primera, de calibración, consiste en determinar los parámetros de las preguntas que son asumidas dentro del modelo. La segunda, de *estimación*, los parámetros de las preguntas estimados de la etapa

[4] Cuando hay solo dos categorías de respuesta y una puede ser considerada mejor que la otra el Modelo de Respuesta Graduada es equivalente a al modelo de TRI dos parámetros propuesto por Birnbaum (1968).

anterior son usados para estimar la medida de interés en cada uno de los evaluados. Es preciso señalar que el proceso de estimación de medidas para cualquier aplicación futura del instrumento estará basado en los resultados de la calibración de las preguntas en la población de calibración considerada en este estudio que definiremos posteriormente.

Para construir una población de calibración de modo que los resultados sean estables en el tiempo y considerando que la muestra del estudio piloto no es representativa de la población correspondiente, se realizaron acciones buscando obtener una muestra de referencia adecuada. Para mejorar la representatividad estadística se optó por utilizar pesos y ponderaciones a través de un proceso de post estratificación en función de los resultados de las pruebas Saber y de las condiciones socioeconómicas de las poblaciones que atienden los colegios. El uso del sistema de ponderaciones fue determinado de modo que la muestra obtenida en el piloto del 2019 se asemejará a una muestra hipotética que fuera representativa de la población escolar de Colombia. Mayores detalles son explicados después. Por otro lado, debido que a la muestra obtenida aún con pesos es solamente una de las posibles muestras que pueden ser obtenidas, se usaron técnicas de remuestreo para mejorar la precisión de la calibración de los parámetros simulando submuestras de la muestra observada en el piloto. Esta población final constituye la población de calibración.

Las medidas aquí obtenidas y reportadas pueden ser usadas para cualquier estudio cuantitativo como análisis de regresión, ecuaciones estructurales, test de comparación, etc, en reemplazo de los tradicionales puntajes fila o puntajes suma que se venían usando. Cabe señalar que, con la metodología adoptada, tendremos medidas en una misma escala de valores que son continuas y resultan comparables a otras medidas en el tiempo y en entre observaciones, las cuales son independientes del número de preguntas incluidas en cada medida y del tipo de medida considerada. Adicionalmente las desviaciones son interpretables como desvíos de la media de referencia adoptada.

En cuanto a la definición de puntos de corte para propósitos de la clasificación de las medidas obtenidas en niveles bienestar, se optó por una postura ecléctica asumiendo la formación de niveles de desempeño usando las medidas TRI basadas en la muestra de referencia usando pesos. Esta construcción de niveles de desempeño usando el puntaje TRI y tomando en consideración la población de calibración para construir niveles de desempeño es simple y aprovecha la interpretabilidad de la escala usada para reportar los resultados.

El presente documento está organizado de la siguiente manera: en el capítulo 1 es definido el estudio psicométrico que consiste en establecer los objetivos del estudio, el marco conceptual y la metodología adoptada. En el capítulo 2 son presentados los resultados del estudio y finalmente en el capítulo son presentadas las Conclusiones y Sugerencias del estudio.

Capítulo 1. Definición del estudio psicométrico

En este capítulo se presentan los objetivos del estudio, el marco conceptual empleado y la metodología.

1.1. Objetivos del estudio

El objetivo central del estudio es aplicar Teoría de Respuesta a Ítem (TRI) al cuestionario de medición de habilidades socioemocionales. Las características psicométricas del instrumento, esto es, análisis satisfactorio de preguntas, confiabilidad, unidimensionalidad y validez ya fueron evaluadas con resultados satisfactorios por lo que no serán objeto del presente estudio (ver Nota técnica, 2020). A continuación, se describen los objetivos:

- Generación de la población de calibración a través de uso de ponderaciones y remuestreo[5].
- Calibración de los parámetros de las preguntas de las escalas haciendo uso del Modelo de Respuesta Graduada (MRG) y el Modelo Logístico de 2 Parámetros (2LP) usando la población de calibración.
- Estimación de medidas a partir de MRG y 2LP usando los parámetros de las preguntas calibradas para las muestras de interés.
- Definición de puntos de corte para el análisis de los puntajes estimados TRI.

El primer punto consiste en obtener la población de calibración y se divide en dos partes: en primer lugar, en generar un sistema de ponderaciones de tal forma que la muestra efectiva se asemeje a una muestra hipotética que fuese representativa de la población escolar del país. En segundo lugar, debido a que la muestra con pesos sigue siendo apenas una realización del proceso generador de muestras, se simulan submuestras tomando como insumo la muestra efectiva con pesos.

El segundo punto consiste en estimar, reportar y describir las características de los parámetros de las preguntas que presentan las escalas al usar el MRG y el 2LP. Como se ha mencionado, para mejorar la representatividad y precisión de estas estimaciones se usan ponderaciones y técnicas de remuestreo. Los parámetros obtenidos en este punto serán utilizados para futuras aplicaciones del instrumento.

El tercer punto tiene como objetivo estimar las medidas o constructos en cada escala para cada uno de los estudiantes en la muestra usando la información de los parámetros de las preguntas

[5] Se debe tener presente la nomenclatura utilizada en cuanto a las diferentes muestras obtenidas en el desarrollo de los objetivos del estudio. Así las cosas, La *muestra observada* se refiere a la muestra bruta obtenida a partir del levantamiento de información. La *muestra efectiva* hace referencia a la muestra obtenida a partir de la *muestra de observada* después de realizar un análisis de valores perdidos (mayores detalles en la sección 1.3.1.). Por último, la *población de calibración* se obtiene a partir de la muestra efectiva, después de aplicar técnicas de resmuestro.

obtenidos en el proceso de calibración, así como establecer la metodología para futuras aplicaciones de los cuestionarios.

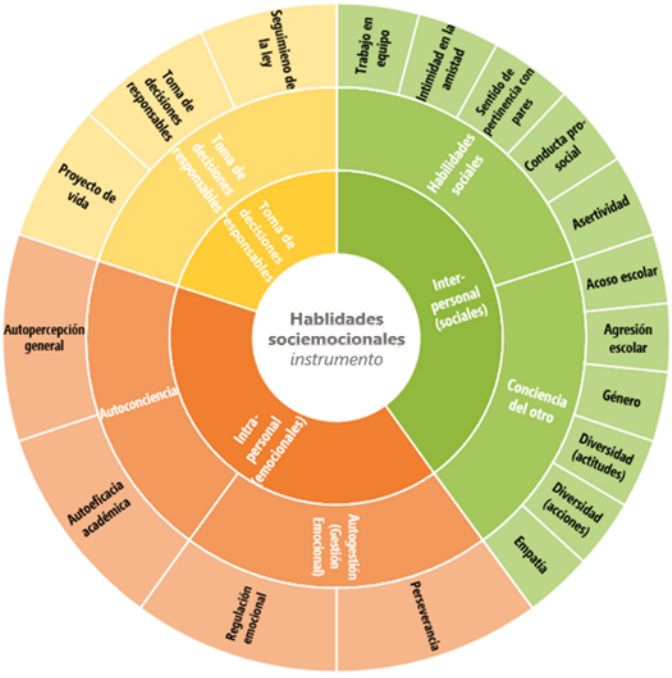
Por último, el cuarto punto presenta una propuesta de criterios de puntos de corte para el análisis de los puntajes estimados bajo MRG y 2PL el cual será considerado para las estimaciones futuras de las medidas en las escalas en nuevas poblaciones de estudiantes.

1.2. Marco conceptual

1.2.1. Instrumento de medición de habilidades socioemocionales

Las escalas del instrumento socioemocional abordan e integran diferentes habilidades, actitudes y comportamientos en el ámbito intrapersonal, interpersonal y cognitivo. A su vez, las escalas se clasifican en las cinco competencias centrales del ampliamente utilizado enfoque conceptual de CASEL (Ilustración 1). En general, el instrumento aborda los procesos a través de los cuales los niños, niñas y jóvenes adquieren y aplican efectivamente los conocimientos, actitudes y habilidades necesarias para identificar y comprender las propias emociones y pensamientos (Autoconciencia), manejar las emociones, establecer y alcanzar metas positivas (Autogestión), sentir empatía por los demás y apreciar la diversidad (Conciencia del otro), establecer y mantener relaciones positivas (Habilidades sociales) y tomar decisiones constructivas y respetuosas (Toma de decisiones responsables).

Ilustración 1: Marco conceptual y habilidades que mide el instrumento



Fuente: elaboración de los autores

La elaboración del instrumento se apoyó en diferentes cuestionarios revisados, adaptados y piloteados en el contexto nacional e internacional. Dentro de las principales fuentes se encuentra el Cuestionario de Bienestar Escolar (CuBe) del Perú y el instrumento de Aulas en Paz[6]. El inventario del CuBE está constituido en un alto porcentaje por adaptaciones de escalas del Middle Years Development Instrument (MDI)[7] y el inventario de Aulas en Paz está conformado principalmente por adaptaciones de escalas que han sido sometidas a evaluación psicométrica en la literatura. Otras fuentes utilizadas son del banco de ítems del ICFES y el instrumento del Programa de Orientación Socio Ocupacional de Fundación Corona. En el Anexo 1 se resume la lista de escalas del instrumento con sus fuentes originales que remite a las evidencias de validez de cada escala.

vLa actual versión del instrumento se deriva del estudio de validación realizado durante el segundo semestre de 2019 y principios de 2020 (ver estudio de validación, 2020). La población de referencia para evaluar las propiedades psicométricas consistió en estudiantes pertenecientes a los grados 4°, 5°, 7°, 8°, 10° y 11° (rangos de edad entre 9 y 18 años), principalmente de instituciones educativas públicas de los departamentos de Caldas, Antioquia, Valle del Cauca y Cesar, y en Bogotá, de instituciones educativas que operan con fondos públicos, pero bajo administración privada. La evaluación dio lugar a un instrumento conformado por tres versiones (una para cada nivel educativo) cuyas escalas poseen altos niveles de confiabilidad y evidencias en favor de su unidimensionalidad. En el Anexo 1 se hallan los resultados de la confiabilidad y la unidimensionalidad de las escalas.

La versión para primaria está conformada por 15 escalas y un total de 70 ítems; la versión para secundaria consta de 16 escalas y 79 ítems, y; la versión para media se conforma de 18 escalas y 90 ítems (Tabla 1). El instrumento emplea ocho distintos formatos de respuesta de tipo politómico y ordinal. En la mayoría de preguntas se usa un formato de 5 opciones: ¡NO!, No, Más o menos, Sí, ¡Sí!. También se emplean dos formatos de frecuencia que tienen 3 opciones: No, A veces, Sí –Ninguna vez, Una vez, 2 o más veces. Además, se usa un formato de cantidad con 4 opciones: Ninguno, Algunos, Muchos, Todos. Otros formatos utilizados son: Feliz, indiferente, triste – No me gustaría, Me da igual, Sí me gustaría – No, sí. La escala de Asertividad posee un formato de respuesta de 6 opciones de tipo situacional, por lo que cada opción de respuesta se ajusta en función la situación planteada en cada item.

Tabla 1: Escalas del instrumento para primaria, secundaria y media

N°	Nombre de la escala	N° categorías respuesta	Primaria			Secundaria			Media		
			Ítems	Ítems Inversos	Total Ítems	Ítems	Ítems Inversos	Total Ítems	Ítems	Ítems Inversos	Total Ítems
1	Acoso escolar	2	53-57	53-57	5	59-63	59-63	5	71-75	71-75	5
2	Agresión escolar	3	45,49-52	45,49-52	5	50,54-58	50,54-58	6	62,66-70	62,66-70	6

[6] Aulas en Paz (PAP) es un programa multicomponente que busca prevenir la agresión y promover formas de convivencia pacífica por medio del desarrollo de competencias ciudadanas en los niños y niñas.

[7] Este instrumento fue elaborado por investigadores de la Universidad de British Columbia en Vancouver, Canadá (Schonert-Reichl, Guhn, Gadermann, Hymel, Sweiss y Hertzman, 2012),

N°	Nombre de la escala	N° categorías respuesta	Primaria			Secundaria			Media		
			Ítems	Ítems Inversos	Total Ítems	Ítems	Ítems Inversos	Total Ítems	Ítems	Ítems Inversos	Total Ítems
3	Asertividad	6	41-44		4	46-49		4	58-61		4
4	Autoeficacia académica	5	4-6		3	7-9		3	12-16		5
5	Autopercepción general	5	-	-		1-3		3	1-3		3
6	Conducta prosocial	3	46-48		3	51-53		3	63-65		3
7	Diversidad (acciones)	4	58-63	58-63	6	64-71	64-71	8	76-83	76-83	8
8	Diversidad (actitudes)	3	64-70		7	71-78		8	84-90		7
9	Empatía	3	34-40		7	39-45	39-45	7	51-57		7
10	Género	5	19-25	19-25	7	22-28	22-28	7	29-35	29-35	7
11	Grit	5	1-3		3	4-6		3	9-11		3
12	Intimidad en la amistad	5	7-9		3	10-12		3	17-19		3
13	Proyecto de vida	5	-	-	-	-	-	-	4-8		5
14	Regulación emocional	3	26-29	26-29	4	29-34	29-34	6	36-42	36,38-42	7
15	Seguimiento de la ley	3	-	-		-	-		47-50	47-50	4
16	Sentido de pertenencia con pares	5	16-18		3	19-21		3	26-28		3
17	Toma de decisiones responsables	3	30-33	30-33	4	35-38	35-38	4	43-46	43-46	4
18	Trabajo en equipo	5	10-15		6	13-18		6	20-25		6

Fuente: elaboración de los autores

En general las preguntas son redactadas en el sentido de que expresan aspectos que denotan mayores habilidades o que favorecen el bienestar de los estudiantes siendo en este caso denominadas preguntas directas o positivas. Sin embargo, existen preguntas, denominadas inversas o negativas, donde por el contrario el acuerdo no denota mayores habilidades o favorece al bienestar. En estos casos el sentido de las opciones de respuesta se codifica en sentido inverso. Estos casos son mostrados en la Tabla 1.

Todas las escalas son codificadas utilizando números enteros consecutivos y siempre asignando el número 0 a la respuesta que se considera peor. Por ejemplo, para el formato de respuesta -¡NO!, No, Más o menos, Sí, ¡Sí! (asumiendo que ¡NO! Denota la peor respuesta y ¡Sí! La mejor), para la categoría de respuesta ¡NO! se asigna un puntaje de 0 mientras que para la categoría de respuesta ¡Sí! se asigna un puntaje de 4.

Adicionalmente el cuestionario contiene preguntas asociadas a características sociodemográficas personales y familiares, posesión de bienes y salud, entre otras conformando la llamada Parte I que no es analizada en este estudio.

1.2.2. El modelo de medición

Para estimar las medidas debemos considerar un marco referencial teórico (modelo de medición) que se adopta en el análisis de los cuestionarios en sus diversas etapas. El modelo de

medición debe ayudar a comprender y a evaluar los puntajes que vienen de las respuestas a las preguntas y de aquí hacia el constructo, y este también debe guiar el uso de los resultados en aplicaciones prácticas. Simplemente, el modelo debe traducir puntajes de respuesta a localizaciones en el mapa del constructo que se quiere medir (Wilson, 2005).

En el ámbito psicométrico existen dos modelos de medición: Teoría Clásica de los Tests (Spearman, 1904; Lord y Novick, 1968) y los modelos de la familia de la Teoría de Respuestas al Ítem o TRI (Baker y Kim, 2004), donde los modelos de la familia de Rasch son casos especiales. En el primer estudio psicométrico (ver Nota técnica, 2020) se utilizó el primer modelo de medición para el análisis de los cuestionarios.

La Teoría Clásica de los Tests es un enfoque según el cual el resultado de la medición de una variable depende de la prueba utilizada y de los sujetos evaluados. El total de los resultados se pueden presentar como el total del puntaje de las preguntas que lo conforman. En TCT la puntuación verdadera predice el nivel de la variable latente y la puntuación observada. Las discrepancias entre el puntaje observado y el puntaje verdadero, es decir el error de medición, se distribuye normalmente con una media de 0 y una desviación estándar de 1.

La teoría de la respuesta al ítem (IRT; Lord, 1952, 1980), también conocida como la teoría del rasgo latente, se refiere a una familia de modelos matemáticos modernos que intentan explicar la relación entre los rasgos latentes (característica o atributo no observable) y sus manifestaciones (es decir, resultados, respuestas o desempeño observado). Estos modelos establecen un vínculo entre las propiedades de los ítems de un instrumento (ej., la facilidad para elegir un ítem que refleja, por ejemplo, autoeficacia), los individuos que responden a estos ítems y el rasgo subyacente que se mide. TRI asume que el constructo latente (ej., nivel de estrés, empatía, depresión) y los ítems de una medida están organizados en un continuo no observable. Por lo tanto, su propósito principal se centra en establecer la posición del individuo en ese continuo (De Ayala, 2009).

La tabla continuación muestra las principales diferencias entre la teoría clásica y teoría de respuesta al ítem.

Tabla 2: Teoría Clásica vs. Teoría de Respuesta al Ítem

Teoría clásica	Teoría de respuesta al ítem
La unidad de análisis es el cuestionario	La unidad de análisis es el ítem
Medidas con más ítems son más confiables	Medidas con menos ítems pueden ser más confiables
La comparación entre diferentes medidas solo puede hacerse cuando los cuestionarios/medidas son paralelas[8]	Las respuestas a los ítems de diferentes medidas pueden ser comparables siempre y cuando estén midiendo el mismo rasgo latente

[8] Las medidas paralelas tienen puntajes verdaderos idénticos y errores linealmente y experimentalmente independientes que tienen varianzas iguales (Lord & Novick, 1968).

Las propiedades de los ítems dependen de una muestra representativa	Las propiedades de los ítems no dependen de una muestra representativa
La posición en el continuo del rasgo latente de un individuo se deriva de comparar los puntajes observados con puntajes observados de un grupo de referencia	La posición en el continuo del rasgo latente de un individuo se deriva de comparar la distancia entre ítems en el continuo del rasgo latente
Todos los ítems de una misma medida deben tener las mismas categorías de respuesta	Los ítems de una misma escala pueden tener diferentes categorías de respuesta

Fuente: Population Health Methods, Columbia

En resumen, el uso de TRI tiene tres grandes ventajas con respecto a la TCT: Primero, permite obtener medidas normales del tipo z o calificación estándar. Segundo, permite establecer parámetros o baremos independientes de la muestra, es decir, que no varían significativamente entre una muestra y otra por más diferentes que sean, pues la normalidad asumida en el modelo de respuesta graduada corresponde a la población "ideal" y no a la muestra de estudio analizada[9]. Con esto se evitan los sesgos de representación estadística que la muestra pueda tener. Y tercero, permite la comparabilidad entre las medidas.

En general, los cuestionarios deben cumplir con un conjunto de requisitos o características psicométricas las cuales es necesario evaluar considerando el modelo de medición adoptado. En el anexo 2 se describe el marco conceptual acerca de las características psicométricas esperadas de un cuestionario que son análisis satisfactorio de preguntas[10], confiabilidad[11], unidimensionalidad[12] y validez[13], Los métodos presentados se apoyan principalmente en los Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) y el Joint Committee on Standards for Educational Evaluation (2003).

El propósito principal de las preguntas de las escalas es servir de base para estimar un conjunto de medidas de los estudiantes que son subyacentes a sus respuestas a dichas preguntas. El conjunto de preguntas del Instrumento son principalmente preguntas de tipo ordinal y de tipo

[9] Como es indicado en las principales referencias de la TRI, la principal ventaja potencial sobre la TCT es la invariancia de los puntajes de la prueba y de las características de las preguntas (ver por ejemplo, Nering y Ostini, 2011). Esto quiere decir a) Invariancia de los parámetros de las preguntas respecto a la muestra que se calcula. Es decir, que los parámetros de la pregunta no cambian aunque las personas que contesten sean distintas, b) Invariancia del parámetro de la medida del sujeto respecto al instrumento utilizado para estimarla. Es decir, que el nivel de medida de interés de la persona no depende del cuestionario utilizado. Para que esta ventaja sea real, se requiere el cumplimiento de los supuestos del modelo de TRI adoptado el cual es detallado en el Anexo 2.

[10] El análisis satisfactorio de las preguntas consiste en identificar y eliminar las preguntas que no satisfacen un conjunto de propiedades que hacen de la prueba un instrumento eficaz y sin sesgo en la estimación de la dimensión medida.

[11] La confiabilidad se refiere a la consistencia de las medidas o puntajes del cuestionario cuando los procedimientos de evaluaciones son repetidos en poblaciones de individuos o grupos admitiendo la presencia de un componente de error.

[12] La unidimensionalidad se refiere a la propiedad de un factor o área de un cuestionario para medir únicamente un constructo o desempeño (unicidad de la prueba medible), esto es, establecer si el conjunto de preguntas dentro de cuestionario mide una sola dimensión. El cumplimiento de esta propiedad es un requisito fundamental para el uso del modelo de Respuesta graduada adoptado en este informe

[13] La validez se refiere al grado por el cual la evidencia y la teoría respaldan las interpretaciones de los puntajes de los cuestionarios.

binario, por lo tanto, en el caso de las ordinales, pueden ser evaluadas siguiendo un modelo de respuesta graduada (Samejima, 1969, Nering y Ostin, 2011) y, en el caso de las binarias, un modelo logístico de 2 parámetros (Birnbaum, 1968). El modelo de respuesta graduada (MRG) y el modelo de dos parámetros (2PL) son modelos TRI apropiado para estos casos pues estiman las medidas subyacentes al conjunto de preguntas consideradas variables latentes. Estas medidas o variables latentes incorporan el "peso" o características de las preguntas en la determinación de dichas variables considerando un determinado modelo probabilístico. De esta manera las variables medidas (escalas o dimensiones) no son una suma simple de los valores de las categorías de respuesta de las preguntas que lo conforman.

Tanto en el modelo de respuesta graduada como en el modelo de dos parámetros (como en los demás modelos de Teoría de Respuesta al Ítem), suponemos que el constructo medido es una variable latente subyacente al conjunto de respuestas dadas a las preguntas que el estudiante responde según el formato de respuesta ordinal o binaria. Un nombre genérico adoptado en estas situaciones es el de constructo o medida. Mayores detalles de estos modelos en el Anexo 3.

1.3. Metodología de estudio

A continuación, se describe la metodología de estudio, indicando la muestra analizada y los procedimientos específicos utilizados en el estudio.

1.3.1. Definición de la muestra efectiva

Para este estudio se considera la muestra recabada a través del piloto para validar el instrumento realizado en 2019 y, además, la muestra proveniente de algunas aplicaciones subsiguientes durante el primer semestre de 2020 realizadas en 22 instituciones educativas públicas ubicadas en su gran mayoría en la ciudad de Santiago de Cali. La muestra piloto-2019 está conformada por 5980 observaciones distribuidas de la siguiente manera: 2023 observaciones en primaria, 2189 observaciones en secundaria y 1769 observaciones en media. La muestra proveniente de aplicaciones en 2020 se distribuye así: 303 observaciones en secundaria y 124 observaciones en media. En total la muestra de referencia está compuesta por 2023 observaciones en primaria, 2462 observaciones en secundaria y 1892 en media.

Los bajos niveles de respuesta a los cuestionarios pueden afectar la validez de la información a usar en el estudio. Así que, con el fin con el fin de depurar y obtener una muestra homogénea y consistente para el análisis psicométrico, solo se consideraron los cuestionarios con un porcentaje de respuesta mayor o igual al 80% (tomando como referencia la parte socioemocional y no la de contexto). Está depuración dio lugar a una muestra efectiva (la utilizada en el presente estudio) conformada por 5785 observaciones distribuidas de la siguiente forma: 1789 observaciones en primaria, 2220 observaciones en secundaria y 1892 observaciones en media.

Con el fin de caracterizar la muestra efectiva (y más adelante definir pesos para mejorar su representatividad con respecto a la población escolar de Colombia) se construyeron estratos en función de los últimos resultados disponibles de las pruebas Saber[14] y el índice de nivel socioeconómico[15] construido por el Icfes. Para el caso de primaria y secundaria se utilizaron los resultados de la prueba en 2017 de lenguaje[16] de quinto y noveno, respectivamente. Para el caso de media se utilizaron los resultados en lenguaje[17] de la prueba aplicada en 2019.

En la siguiente tabla se presenta la distribución de la muestra efectiva para los tres niveles educativos en función de 16 estratos construidos a partir de los cuartiles de los resultados en la prueba estandarizada y los cuartiles del índice socioeconómico, comparando con la distribución de la población en estos estratos. Los valores de la información poblacional corresponden al número de evaluados en las pruebas censales.

Tabla 3: Distribución de la muestra efectiva de estudiantes en estratos según cuartiles del nivel socioeconómico y resultados en pruebas Saber de Lenguaje

Cuartiles desempeño Lenguaje[1]	Cuartiles NSE[2]	Primaria				Secundaria				Media			
		Población		Muestra		Población		Muestra		Población		Muestra	
		N°	%	N°	%	N°	%	N°	%	N°	%	N°	%
Q1	Q1	50157	6%	40	2%	34749	6%	103	5%	23996	7%	56	3%
	Q2	120228	15%	398	22%	57431	10%	227	10%	34887	10%	343	19%
	Q3	46832	6%	531	30%	16508	3%	78	4%	8543	2%	12	1%
	Q4	1209	0%	-	0%	485	0%	-	0%	410	0%	-	0%
Q2	Q1	30670	4%	-	0%	21289	4%	62	3%	16467	5%	188	11%
	Q2	88892	11%	-	0%	59039	10%	231	10%	40158	11%	335	19%
	Q3	125393	16%	215	12%	65761	11%	711	32%	36146	10%	178	10%
	Q4	8417	1%	-	0%	3114	1%	-	0%	3203	1%	-	0%
Q3	Q1	17106	2%	-	0%	8682	1%	27	1%	6462	2%	30	2%
	Q2	36447	5%	-	0%	32791	6%	-	0%	23253	6%	24	1%

[14] Las pruebas Saber son evaluaciones externas estandarizadas aplicadas por el Icfes, las cuales evalúan el desempeño alcanzado por los estudiantes según las competencias básicas definidas por el Ministerio de Educación Nacional. Estas pruebas evalúan los desempeños desarrollados por los estudiantes al final de los ciclos de los niveles educativos de la educación básica y media. Saber 3º y 5º en la básica primaria, Saber 9º en el cierre de la educación básica secundaria, y Saber 11º al término de la educación media.

[15] El índice socioeconómico es calculado por el Icfes a partir de la información recolectada en el cuestionario de contexto que se aplica en conjunto con las pruebas Saber. Para la construcción del índice se utilizan variables relacionadas con la educación de los padres, las condiciones habitacionales del hogar, el acceso a servicios públicos y la posesión de bienes. Además, incluye un componente cultural que abarca aspectos como la asistencia a actividades de teatro, parques, bibliotecas, entre otros.

[16] Las pruebas de quinto y noveno aplicadas por el Icfes en 2017 (última aplicación censal para estos grados) evalúan competencias en lenguaje y matemáticas. Se escogió lenguaje debido a la que mayor cobertura de la población con respecto a la muestra del piloto. Sin embargo, los resultados de ambas pruebas están altamente correlacionados por lo que las conclusiones no cambiarían significativamente si se utilizarán los resultados de matemáticas.

[17] En media se escogió lenguaje para mantener la consistencia con la elección de prueba hecha en primaria y secundaria.

Cuartiles desempeño Lenguaje[1]	Cuartiles NSE[2]	Primaria				Secundaria				Media			
		Población		Muestra		Población		Muestra		Población		Muestra	
		N°	%	N°	%	N°	%	N°	%	N°	%	N°	%
	Q3	99233	13%	274	15%	123877	21%	273	12%	68923	19%	191	11%
	Q4	37251	5%	-	0%	19447	3%	-	0%	13275	4%	38	2%
Q4	Q1	3523	0%	-	0%	1628	0%	-	0%	820	0%	-	0%
	Q2	5077	1%	-	0%	2898	0%	-	0%	3382	1%	-	0%
	Q3	20438	3%	81	5%	37931	6%	282	13%	23492	6%	294	17%
	Q4	85291	11%	250	14%	103758	18%	226	10%	60694	17%	87	5%
Total		776164	100%	1789	100%	589388	100%	2220	100%	364111	100%	1776	100%

[1] Cuartiles de la distribución de la distribución de resultados en lenguaje de las pruebas Saber.
[2] Cuartiles de la distribución del índice socioeconómico del Icfes.
[3] Las celdas resaltadas en verde hacen referencias a los estratos en dónde hay observaciones de la muestra efectiva.
Fuente: Elaboración de los autores con base en resultados pruebas Sabes (Icfes) y la muestra de estudio.

De acuerdo con la Tabla 3 notamos que la muestra en los tres niveles educativos no corresponde a la distribución esperada siguiendo los datos poblacionales (lo que se evidencia al ver tanto los estratos cubiertos como las participaciones dentro de cada estrato). Especialmente este fenómeno es más marcado en el caso de primaria en dónde solo hay muestra para 7 de los 16 estratos. No obstante, en el caso de secundaria y media hay observaciones en la muestra efectiva en 10 y 12 estratos respectivamente y, además, los estratos que no están representados en la muestra efectiva resultan ser los que menor población abarcan. En general, la muestra del estudio cuenta con observaciones en estratos que abarcan el 70.6%, 89.8% y 97.85% de la población de primaria, secundaria y media correspondientemente.

Ahora bien, la determinación de las características que presentan las preguntas, así como el proceso de estimación de las medidas se hará a través de modelos TRI que, al contrario de la metodología clásica en que se suman las respuestas de las preguntas que componen la escala, no dependen de la representatividad estadística[18]. No obstante, a pesar de haber sido ampliamente estudiada la propiedad invarianza de los modelos TRI (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991), esta no siempre se cumple en las aplicaciones empíricas (Cook, Eignor, & Taft, 1988; Miller & Linn, 1988). Así las cosas, con el fin de mejorar la representatividad y la precisión de las estimaciones de parámetros y medidas se ha optado por utilizar, por un lado, pesos de tal forma que la muestra efectiva se asimile lo más posible a la población escolar de Colombia como se muestra en la Tabla 3 (ver Anexo 5), y

[18] Como es indicado en las principales referencias de la TRI, la principal ventaja potencial sobre la TCT es la invariancia de los puntajes de la prueba y de las características de las preguntas (ver por ejemplo, Nering y Ostini, 2011). Esto quiere decir a) Invariancia de los parámetros de las preguntas respecto a la muestra que se calcula. Es decir, que los parámetros de la pregunta no cambian aunque las personas que contesten sean distintas, b) Invariancia del parámetro de la medida del sujeto respecto al instrumento utilizado para estimarla. Es decir, que el nivel de medida de interés de la persona no depende del cuestionario utilizado. Para que esta ventaja sea real, se requiere el cumplimiento de los supuestos del modelo de TRI adoptado el cual es detallado en el Anexo 2.

por otro lado, técnicas de remuestreo para garantizar la robustez de los resultados de futuras aplicaciones del Instrumento (más adelante, en la sección 2.1., se describe este proceso en mayor detalle).

En la tabla a continuación se describe el número de estudiantes que respondieron los cuestionarios según diferentes características de interés para los tres niveles educativos. El número total de casos corresponde las respuestas válidas.

Tabla 4: Estadísticas descriptivas muestra efectiva

Variable	Categoría de respuesta	Primaria		Secundaria		Media	
		%	N°	%	N°	%	N°
Genero	Mujer	51%	906	53%	1160	58%	1023
Número de hermanos	0	2%	37	1%	29	2%	27
	1	29%	478	27%	564	23%	393
	2	25%	418	25%	524	27%	457
	3	16%	265	17%	359	18%	303
	4	28%	463	30%	630	31%	523
Estudiantes para los que leer es	Difícil o muy difícil	8%	143	7%	141	8%	135
Estudiantes extranjeros	Venezolano(a) u otra nacionalidad	1%	23	1%	33	1%	10
Origen étnico	Indígena	3%	51	3%	65	2%	40
	Blanco o mestizo	51%	894	55%	1198	56%	982
	Negro, afrodescendiente o afrocolombiano	17%	294	23%	494	26%	460
	Otro	4%	68	3%	65	3%	47
	No sabe	26%	450	17%	367	13%	226
Estudiantes que son acompañados por un adulto al colegio		60%	1039	33%	707	19%	337
Estudiantes con acceso a internet en casa		71%	1242	71%	1558	67%	1174
Estado de salud	Mala	1%	15	1%	25	2%	36
	Regular	15%	260	18%	395	23%	405
	Buena	38%	672	48%	1059	55%	965
	Excelente	46%	822	33%	731	21%	364
Estudiantes con algún tipo de incapacidad	Por ejemplo, sordera, cojera, uso de silla de ruedas	8%	139	4%	87	4%	64
Nivel educativo padre	Ninguno	4%	62	3%	73	4%	75
	Primaria	13%	211	18%	389	26%	448
	Secundaria	7%	110	13%	274	12%	198
	Media	19%	312	20%	428	21%	364
	Superior	17%	272	20%	437	22%	371
	No sabe	40%	656	25%	542	15%	253
Nivel educativo madre	Ninguno	3%	43	2%	41	3%	51
	Primaria	13%	215	17%	373	22%	377
	Secundaria	9%	152	14%	301	13%	227
	Media	25%	419	28%	614	28%	498
	Superior	19%	315	24%	529	27%	475
	No sabe	31%	509	14%	313	7%	122
Tenencia de lavadora	Tiene lavadora	91%	1589	89%	1932	85%	1485
Tenencia de computados	Sin computador	36%	605	31%	656	37%	631
	Un computador	48%	802	47%	990	48%	815
	Dos computadores	16%	269	23%	480	15%	262
Tenencia de televisor	Sin televisor	4%	63	8%	168	4%	68
	Un televisor	49%	852	52%	1055	56%	953
	Dos televisores	47%	820	40%	820	39%	666
Tenencia de carro	Sin carro	69%	1083	61%	1263	78%	1303
	Un televisor	25%	397	20%	412	18%	301
	Dos televisores	6%	99	19%	393	4%	62
~~Tenencia de Moto~~	Sin moto	44%	728	43%	892	53%	892

Variable	Categoría de respuesta	Primaria		Secundaria		Media	
		%	N°	%	N°	%	N°
	Una moto	43%	709	43%	890	39%	652
	Dos motos	13%	220	15%	312	9%	148

Fuente: elaboración de los autores

1.3.2. Plan de implementación del estudio

A continuación, se presenta la secuencia considerada para el análisis considerado en este estudio.

a) Construcción de pesos y ponderaciones y uso de remuestreo para definir la población de calibración

El primer objetivo es, a través de ponderaciones, construir una muestra a partir de la muestra efectiva que se asemeje a una muestra que fuese representativa de la población escolar del país. La construcción de pesos se realizará por medio de un proceso de posestratificación en función de los resultados en pruebas estandarizadas y el nivel socioeconómico de las poblaciones que atienden los colegios (ver tabla Tabla 3). Además, a partir de la muestra con pesos se van a definir la población de calibración en base a la selección de muestras aleatorias bajo un muestreo aleatorio simple (sin reposición) considerando un tamaño de muestra para poblaciones finitas usando el muestreo aleatorio simple. El conjunto de muestras de calibración derivado de este proceso constituirá la base para el cálculo de los parámetros de los ítems del instrumento y es considerado la población de calibración o población de referencia que es una aproximación de la verdadera población sobre la que medimos las escalas del instrumento, que es en la práctica desconocida.

b) Aplicación del modelo de respuesta graduada para estimar los parámetros de las preguntas de las escalas.

El objetivo es reportar las características de los parámetros de preguntas que presentan las escalas cuando es asumido el modelo de respuesta graduada (en el caso es escalas con formato de respuesta politómico) el cual coincide con el modelo logístico de dos parámetros (en el caso de escalas con formato de respuesta binario). Para esto se estiman estas características siguiendo un proceso de máxima verosimilitud marginal (mayores detalles en el Anexo 3). Este proceso es denominado usualmente etapa de calibración de preguntas. En esta esta etapa se utilizan las muestras de calibración obtenidas en el punto anterior, en dónde se utilizan pesos y remuestreo. Por tanto, a partir de la aplicación de los modelos TRI a las muestras aleatorias (obtenidas de la muestra efectiva con pesos) se aproximará la distribución muestral de los parámetros asociados a las preguntas (Mayores detalles en el Anexo 6).

c) Estimación de las medidas o constructos de las escalas en grupos de interés, usando el modelo de respuesta graduada

El objetivo es estimar las medidas o constructos, en cada escala y nivel educativo, para cada uno de los estudiantes en cualquier muestra de interés. Puede ser aplicado en la muestra efectiva, en nuevas muestras donde el instrumento será aplicado o en estudiantes que responden a las preguntas del instrumento. Se usa las respuestas de los grupos de interés y se usa la información de los parámetros de las preguntas obtenidos en el proceso de calibración y el resultado son las medidas de las escalas del instrumento usando una escala de medida TRI. El proceso permite establecer la metodología para futuras aplicaciones de los cuestionarios. Este proceso es denominado usualmente etapa de estimación de las medidas. Este proceso toma en cuenta los análisis previos y se enfoca en las medidas que satisfacen el cumplimiento de los supuestos de los modelos TRI. Las medidas obtenidas son denominadas medidas TRI y se presentan en una escala con media 50 y desviación estandar 10 y se espera que el intervalo entre 30 y 70 contenga aproximadamente 95% de las veces las medidas TRI de las muestras de interés.

d) Determinación de puntos de corte para interpretación de los constructos estimados para las escalas.

El objetivo es definir una metodología para establecer puntos de corte[19] que permiten transformar las medidas TRI de las escalas del instrumento en categorías o niveles de desarrollo interpretables. Los cuales serán usados para cada uno de los estudiantes en cualquier muestra de interés, incluyendo la muestra efectiva. Este proceso puede ser usado para establecer la metodología para futuras aplicaciones de los cuestionarios y es denominado generalmente etapa de interpretación de las medidas.

Este proceso es usualmente complejo y requiere metodologías específicas para su elaboración[20]. Los métodos pueden ser clasificados en dos grupos principales, aquellos que son determinados a priori y aquellos que son elaborados a posteriori, estos últimos son especialmente basados en el uso de los resultados de la Teoría de Respuesta al Ítem. En todos los casos sin embargo se requiere de la participación de un conjunto de especialistas que participan en diferentes momentos de la metodología adoptada para interpretar adecuadamente las medidas. Esto es así porque al adoptar la metodología de TRI se está frente a un modelo referido a criterios el

[19] El proceso práctico para determinar los puntajes de corte (cutoff score en inglés) es denominado comúnmente como la definición de estándares.

[20] Como indican Crocker y Algina (2006), en términos generales hay tres enfoques para la definición de estándares. La primera consiste en una inspección del contenido de una prueba por uno o más jueces expertos que emitir un juicio sobre la base de una impresión global de contenido de la prueba, y el segundo está basado en los juicios de contenido de preguntas individuales y el tercero se basa en el desempeño de los examinados y aunque es algo más psicométrica que las anteriores también implica un elemento importante de juicio informado ya que todos los métodos clasificados bajo este enfoque requieren definir el estándar para elegir los grupos del examinando cuyo desempeño se examina.

cual difiere sustancialmente del más conocido modelo normativo que está asociado a la metodología TCT[21] (Ver más detalles en el Anexo 2).

En el primer estudio piloto se utilizó una aproximación a priori que consistió en una inspección del contenido de los cuestionarios por parte de un juez experto que emitió un juicio sobre el contenido de las preguntas del cuestionario, dando lugar así a un punto de corte para cada escala (es decir dos niveles de clasificación). Los puntos de corte obtenidos bajo está metodología generaron resultados con muy baja varianza que no permitían discriminar estudiantes ni colegios de acuerdo con los resultados obtenidos en los cuestionarios. Estos resultados tampoco parecían ser coherentes cuando se confrontaban con la experiencia y el conocimiento que los usuarios del instrumento tenían sobre las instituciones educativas. También, existía dificultad en la comparación entre resultados de diferentes medidas y entre niveles. Cabe mencionar que los resultados utilizando los puntajes brutos (construidos a partir de Teoría Clásica de los Tests) si presentaban varianza y en un análisis preliminar de factores asociados se encontraron hallazgos en línea con la literatura (poblaciones en desventaja como estudiantes en extraedad, en condición de repitencia o discapacidad obtienen significativamente peores resultados) pero la dificultad estuvo en la creación de niveles de desarrollo de las medidas que presenten interpretabilidad.

Para solucionar los problemas arriba descritos, en este estudio se aplica un enfoque a posteriori en dónde las medidas obtenidas a través de la aplicación de TRI se utilizan en la elaboración de los criterios de clasificación. Como se verá más adelante, el uso TRI resulta conveniente ya que permite aprovechar que las medidas del instrumento son asumidas normalmente en la población de referencia y por lo tanto no es necesario realizar un proceso de transformación para normalizarlas. Así las cosas, es posible hacer uso directo de las calificaciones estándar para estimar los puntos de corte.

Capítulo 2. Resultados

1.1. Construcción de pesos y ponderaciones y uso de remuestreo para definir la población de calibración

Como se ha mencionado en repetidas ocasiones a lo largo del documento, el uso de ponderaciones obedece al hecho de que la muestra disponible para este estudio no corresponde necesariamente a una muestra representativa de la poblacional escolar de Colombia. Para ilustrar lo anterior, las tablas a continuación (ver Tabla 4, Tabla 5 y Tabla 6),

[21] En otras palabras, mientras el modelo normativo construye sus puntos de corte para la interpretación de las medidas considerando una muestra normativa (representativa de la población de interés) estableciendo baremos que serán usados para aplicaciones futuras, los modelos referidos a criterios caracterizan las medidas independientes de la muestra de estudio, estableciendo juicios de equivalencias entre los resultados estimados y esperados para determinar puntos de corte definidos.

comparan la distribución de las muestras efectivas de primaria, secundaria y media con la distribución de la población escolar del país en cada nivel en función de los resultados en pruebas estandarizadas en lenguaje y el nivel socioeconómico de estas poblaciones (ver sección 1.3.1 sobre la definición de la muestra efectiva para mayor detalle).

En general, como fue notado antes, a pesar de que la muestra efectiva parece ser bastante diversa, esta no cubre todos los estratos presentes en la población escolar de referencia, e incluso en los estratos cubiertos, las participaciones tienden a diferir también con aquellas encontradas en la población de referencia. Por tanto, la definición de pesos busca que la muestra efectiva se asemeje a la población escolar, dando mayor peso a aquellas observaciones que están subrepresentadas y dando menor peso a las observaciones que se están sobrerrepresentadas en la muestra observada.

Tabla 5: Cálculo de ponderaciones para muestra de primaria

Cuartiles desempeño Lenguaje[1]	Cuartiles NSE[2]	Primaria				Participación nacional estratos representados en muestra	Pesos
		Población		Muestra			
		N°	%	N°	%		
Q1	Q1	50157	6%	40	2%	9%	4.10
	Q2	120228	15%	398	22%	22%	0.99
	Q3	46832	6%	531	30%	9%	0.29
	Q4	1209	0%	-	0%	-	
Q2	Q1	30670	4%	0	0%	-	-
	Q2	88892	11%	0	0%	-	-
	Q3	125393	16%	215	12%	23%	1.91
	Q4	8417	1%	-	0%	-	-
Q3	Q1	17106	2%	0	0%	-	-
	Q2	36447	5%	0	0%	-	-
	Q3	99233	13%	274	15%	18%	1.18
	Q4	37251	5%	0	0%	-	-
Q4	Q1	3523	0%	-	0%	-	-
	Q2	5077	1%	-	0%	-	-
	Q3	20438	3%	81	5%	4%	0.82
	Q4	85291	11%	250	14%	16%	1.11
Total		776164	100%	1789	100%	NA	NA

[1] Cuartiles de la distribución de la distribución de resultados en lenguaje de las pruebas Saber.
[2] Cuartiles de la distribución del índice socioeconómico del Icfes.
[3] Las celdas resaltadas en verde hacen referencias a los estratos en dónde hay observaciones de la muestra efectiva.
Fuente: Elaboración de los autores con base en resultados pruebas Sabes (Icfes) y la muestra de estudio.

A modo de ejemplo para ilustrar la generación del sistema de ponderaciones, obsérvese la Tabla 4, en dónde se presenta el caso para primaria. Como primer paso es importante decir que la definición de pesos se hace sobre los estratos en dónde hay al menos una observación, en este caso hay 9 estratos que no están representados en la muestra observada y, por ende, tampoco estarán representados en la muestra que se deriva de la aplicación de pesos (no obstante, en el caso de primaria, la muestra con pesos será representativa del 70% de la población en primaria. Ver sección 1.3.1).

En segundo lugar, los pesos se pueden entender como el factor por el cual se debe expandir (o contraer) cada observación de la muestra efectiva con el fin de que la distribución de la muestra en cada estrato se asemeje a la distribución esperada según la muestra de referencia. Así las cosas, los valores de la columna "Pesos" se calculan al dividir el porcentaje de la población de referencia en cada estrato (descontando la población en estratos no representados en la muestra efectiva) entre el porcentaje de la muestra efectiva que se encuentra en un estrato determinado (siempre y cuando sea siempre mayor a cero)[22].

Para dar un ejemplo de la interpretación de los pesos tomemos como referencia la muestra efectiva ubicada en el primer estrato para secundaria (conformado por la población de estudiantes de colegios ubicados en el primer cuartil de los resultados en lenguaje y en el primer cuartil del nivel socioeconómico). Aquí el peso quiere decir que cada observación de la muestra efectiva en este estrato contará por 1.42 observaciones, de tal modo que esta población observada no represente el 5% de la muestra sino el 7%.

Tabla 6: Cálculo de ponderaciones para muestra de secundaria

Cuartiles desempeño Lenguaje[1]	Cuartiles NSE[2]	Secundaria				Participación nacional estratos representados en muestra	Pesos
		Población		Muestra			
		N°	%	N°	%		
Q1	Q1	34749	6%	103	5%	7%	1.42
	Q2	57431	10%	227	10%	11%	1.06
	Q3	16508	3%	78	4%	3%	0.89
	Q4	485	0%	-	0%	-	-
Q2	Q1	21289	4%	62	3%	4%	1.44
	Q2	59039	10%	231	10%	11%	1.07
	Q3	65761	11%	711	32%	12%	0.39
	Q4	3114	1%	-	0%	-	
Q3	Q1	8682	1%	27	1%	2%	1.35
	Q2	32791	6%	-	0%	-	-
	Q3	123877	21%	273	12%	23%	1.90
	Q4	19447	3%	-	0%	-	-
Q4	Q1	1628	0%	-	0%	-	-
	Q2	2898	0%	-	0%	-	-
	Q3	37931	6%	282	13%	7%	0.56
	Q4	103758	18%	226	10%	20%	1.93
Total		589388	100%	2220	100%	NA	NA

[1] Cuartiles de la distribución de la distribución de resultados en lenguaje de las pruebas Saber.
[2] Cuartiles de la distribución del índice socioeconómico del Icfes.
[3] Las celdas resaltadas en verde hacen referencias a los estratos en dónde hay observaciones de la muestra efectiva.

[22] Por ejemplo, el peso para el estrato conformado por el primer cuartilobsérvese la primera fila de la tabla para secundaria (Tabla 5),

Fuente: Elaboración de los autores con base en resultados pruebas Sabes (Icfes) y la muestra de estudio.

De otra parte, la muestra efectiva considerando los pesos constituye una población de referencia para la calibración, pero esta no necesariamente es una población generalizable a todas las muestras que se pueden obtener, así que se va a definir una posible población de calibración a partir de la población de referencia. Estas poblaciones de calibración se basan en la selección de una muestra aleatoria bajo un muestreo aleatorio simple (sin reposición) considerando un tamaño de muestra aleatorio simple. Adicionalmente, este proceso es repetido un número de veces generando un conjunto de muestras de calibración, de modo que este conjunto constituye la población de calibración y es la base para el cálculo de los parámetros de los ítems del instrumento. Así, el modelo de respuesta gradual es aplicado para cada población de calibración y para definir los parámetros de calibración de los ítems, se usa la media de las estimaciones de los parámetros en cada población de calibración. Esta estrategia garantiza la robustez de los resultados futuros usando la población de referencia.

Tabla 7: Cálculo de ponderaciones para muestra de media

Cuartiles desempeño Lenguaje[1]	Cuartiles NSE[2]	Media				Participación nacional estratos representados en muestra	Pesos
		Población		Muestra			
		N°	%	N°	%		
Q1	Q1	23996	7%	56	3%	3%	2.96
	Q2	34887	10%	343	19%	19%	0.70
	Q3	8543	2%	12	1%	1%	4.92
	Q4	410	0%	-	0%	-	-
Q2	Q1	16467	5%	188	11%	11%	0.61
	Q2	40158	11%	335	19%	19%	0.83
	Q3	36146	10%	178	10%	10%	1.40
	Q4	3203	1%	-	0%	-	-
Q3	Q1	6462	2%	30	2%	2%	1.49
	Q2	23253	6%	24	1%	1%	6.70
	Q3	68923	19%	191	11%	11%	2.49
	Q4	13275	4%	38	2%	2%	2.42
Q4	Q1	820	0%	-	0%	-	-
	Q2	3382	1%	-	0%	-	-
	Q3	23492	6%	294	17%	17%	0.55
	Q4	60694	17%	87	5%	5%	4.82
Total		364111	100%	1776	100%	NA	NA

[1] Cuartiles de la distribución de la distribución de resultados en lenguaje de las pruebas Saber.
[2] Cuartiles de la distribución del índice socioeconómico del Icfes.
[3] Las celdas resaltadas en verde hacen referencias a los estratos en dónde hay observaciones de la muestra efectiva.
Fuente: Elaboración de los autores con base en resultados pruebas Sabes (Icfes) y la muestra de estudio.

1.2. Aplicación del modelo de respuesta para calibrar los parámetros de las escalas

1.2.1. Estudio de Calibración

Este proceso consiste en determinar los parámetros de las preguntas que son asumidas dentro del modelo de respuesta graduada. Este proceso es llamado proceso de calibración y constituye una forma de verificación definitiva para adoptar el modelo de logístico de dos parámetros y el modelo de respuesta graduada y obtener las ventajas de estos modelos ya citados anteriormente.

Dado que el proceso seguido en este estudio estima las características de las preguntas dentro de cada medida (llamadas parámetros de ítems en la terminología del modelo los modelos TRI), estos valores serán considerados para estimaciones posteriores de las medidas individuales como se discute en la siguiente sección.

Los parámetros estimados de las preguntas usando TRI son de dos tipos (de discriminación y de dificultad o umbral (ver Anexo 3 para detalles). Para una pregunta con cinco categorías de respuesta son definidos un parámetro de discriminación y cuatro parámetros de umbral o dificultad de paso entre categorías (uno menos que el número de categorías). Para una pregunta con tres categorías de respuesta son definidos un parámetro de discriminación y dos parámetros de umbral.

Un parámetro de discriminación representa la capacidad del ítem para discriminar en la medida considerada. Este parámetro toma valores positivos, pero se espera que no sean muy altos o pequeños. Así son preferibles preguntas con discriminación alrededor de 2, siendo valores entre 1 y 2 aceptables (ver detalles en el Anexo 3).

Dado que estamos considerando respuestas en sentido positivo (pasando de categorías de respuesta ordinal ¡No! hasta ¡Sí! o pasando a la mayor ocurrencia en las respuestas de frecuencia),
esperamos que sea más difícil concordar con las conductas descritas que sean más positivas. Esto se refleja en el parámetro de dificultad del modelo. En este caso este va de valores negativos hasta positivos aproximadamente entre -5 y 5, es decir con un rango de valores equivalente al de la medida. Este parámetro representa la "facilidad" (cuando es más negativo) o "dificultad (cuando es más positivo) que tiene pasar de una categoría de respuesta a la siguiente en el orden de valores de respuesta.

Perseverancia									
Pregunta	Media (mean)	Desviación estándar (sd)	Correlación item-total resto de items de la escala (corr)	Carga factorial en la dimensión de la	Discriminación del ítem (a)	Dificultad de paso entre la 1ª y 2ª categoría (b1)	Dificultad de paso entre la 2ª y 3ª categoría (b2)	Dificultad de paso entre la 3ª y 4ª categoría (b3)	Dificultad de paso entre la 4ª y 5ª categoría (b4)

				escala (load)					
p7	3.05	1.01	0.62	0.55	1.849	-2.639	-2.099	-0.984	0.309

A modo de ilustración mostramos el caso de la primera pregunta para escala de Perseverancia en la versión para primaria que tiene 5 categorías de respuesta ¡NO!, No, Más o menos, Sí, ¡Sí!. Considerando el modelo de respuesta graduada notamos que presenta un parámetro de discriminación (a) de 1.849 que indica que es una pregunta adecuada para medir la Perseverancia. Los resultados de los umbrales o dificultades de paso (b) indican en los 3 primeros casos valores negativos que expresan la relativa facilidad en concordar con las primeras 4 categorías de respuesta, inclusive entre el paso del *Más o menos* a *Sí* donde encontramos un valor de -0.984. Por otro lado, el parámetro b4 es positivo e indica el grado de dificultad en el paso entre *Sí* a *¡Sí!* y señala que fue relativamente difícil concordar con la conducta descrita en esta pregunta y categoría de respuesta. La tabla anterior también reporta resultados presentados en el primer estudio considerando el modelo TCT. En la primera y segunda columna son mostradas estadísticas descriptivas (media y variancia) del ítem. En la tercera columna, la correlación ítem total que indica la coherencia interna del ítem al medir lo mismo que otros ítems en la escala, y finalmente en la cuarta columna mostramos la carga factorial, que en este caso, siendo superior a 0.30 indica que este ítem está adecuadamente representado en la escala.

1.2.2. Resultados del estudio de calibración

Todos los resultados del estudio de calibración de las preguntas ajustando el modelo respuesta graduada para cada una de las escalas en primaria, secundaria y media del Instrumento son presentadas en el Anexo 4.

Aunque no existe un criterio definitivo para evaluar la calidad de ajuste de los los parámetros usando el modelo de respuesta graduada, Zickar, Russel, Smith, Bohle y Tilley (2002) sugieren que todos los parámetros de discriminación mayores que 1 indican una discriminación aceptable entre personas. Así las cosas, en este estudio seguiremos sus pautas de tal manera que se considera que valores del parámetro de discriminación menores a 1 denotan ítems de baja calidad, valores entre 1 y 2 indican calidad moderada y valores superiores a 2 indicar alta calidad.

Así las cosas, en primaria, de las 70 preguntas, ninguna tiene un parámetro de discriminación menor a 1, 34 preguntas tienen parámetros de discriminación con valor entre 2 y 3, y 36 parámetros con discriminación mayor a 1. En secundaria, de las 78 preguntas, ninguna tiene discriminación menor a 1, 24 tienen discriminación entre 1 y 2, y 54 preguntas tienen discriminación mayor a 2. En media, de las 90 preguntas, tan solo 4 tienen discriminación

menor a 1 (pero no menor a 0.92), 27 preguntas tienen discriminación entre 1 y 2, y 58 tienen discriminación mayor a 2.

En resumen, los resultados del estudio de calibración de las preguntas ajustando el modelo de respuesta graduada para las escalas de primaria, secundaria y media muestran que las preguntas presentan parámetros adecuados que indican que todas las escalas presentan un buen ajuste usando el modelo. Esto significa que para futuras aplicaciones del Instrumento los resultados de la calibración, es decir, los parámetros de las preguntas (dificultad de paso entre categorías o umbrales y discriminación) podrán ser asumidos como fijos o conocidos. Adicionalmente, desde que estos parámetros de calibración no fueron calculados con la muestra efectiva de este estudio y si con una población de calibración, estos resultados son asumidos como siendo resultados obtenidos en una población de referencia de las escalas medidas y son robustos en el tiempo y pueden ser usados para diferentes muestras de interés, los cuales serán ''calibradas'' en relación a esta población de referencia. Eventualmente, la calibración puede ser revisada futuramente, de existir cambios importantes observados en la población.

En primaria, cabe destacar los altos valores para los parámetros de discriminación de algunas preguntas, es decir que discriminan mejor entre encuestados con diversos niveles de determinada habilidad. Es el caso de la pregunta de *"Cuando veo que molestan a un niño que me cae mal, yo me siento"* de la escala de *Empatía;* la pregunta *"Tengo un(a) amigo(a) a quien le puedo contar todo"* de la escala de *Intimidad en la amistad, y; "Las mujeres deben ocuparse de limpiar y cocinar para los hombres"* de la escala de *Genero.*

En secundaria, dentro de las preguntas con alta discriminación, además de las ya descritas en primaria, se puede mencionar: *"Solo las mujeres deben estar en la cocina"* de la escala de *Genero*; *"Tengo al menos un(a) muy buen(a) amigo(a) con quien puedo hablar cuando algo me molesta"* de la escala de *Intimidad en la amistad, y; "¿Alguien de tu salón envía mensajes o fotos desagradables sobre ti a otros TODO EL TIEMPO, por celular o por Internet, haciéndote sentir muy mal?"* de *Acoso Escolar; "Siempre me dedico a algo hasta terminarlo."* De *Perseverancia.*

En Media, además de las ya mencionadas en primaria y secundaria, se pueden destacar las siguientes preguntas con alta discriminación: "*Consolaste a alguien que estaba triste" de la escala Conducta prosocial; "Personas que vienen de otra región del país"* de la escala *Diversidad – actitudes; "¿Alguien de tu salón te saca de los grupos TODO EL TIEMPO, haciéndote sentir muy mal?"* de la escala de *Acoso Escolar; "Llorar es de niñas"* de la escala de *Genero; "Rechazan a otras personas por ser o parecer de una cultura indígena"* de la escala *Diversidad – acciones, y; "Me siento capaz de graduarme del colegio"* de la escala *Autoeficacia académica.*

1.3. Estimación de las medidas para las escalas

Considerando los resultados obtenidos para la estimación de los parámetros de las preguntas o proceso de calibración de la sección anterior, la siguiente etapa es obtener las medidas de las personas en los constructos del Instrumento. Este proceso en la literatura de TRI es llamado proceso de estimación de las medidas y es aplicado para cada una de las personas que son parte de las diferentes muestras que sean de interés. En este caso los parámetros de las preguntas estimados de la etapa anterior son usados para estimar las medidas de los evaluados.

Esto significa que el proceso de estimación (de medidas) para cualquier aplicación futura del Instrumento estará basado en los resultados de la calibración de las preguntas de la muestra considerada en este estudio. Como se ha citado antes, considerando las características del modelo TRI, el proceso de calibración es independiente del proceso de estimación. Las muestras en ambos procesos pueden ser las mismas o pueden ser diferentes. Esto significa que se puede usar la calibración de las preguntas mostradas en el Anexo 4 para cualquier estimación de medidas en cualquier muestra (actual o futura). Este proceso es un proceso computacional que asocia a cada individuo en la muestra una medida (o posición relativa) en cada constructo (escala y dimensión) usando la información de las preguntas reportada en este informe.

La estimación sigue un procedimiento de estimación estadística considerando el modelo de respuesta gradual adoptado en el que es posible usar diferentes métodos de estimación. En este trabajo se adopta el método de estimación Expected A Posteriori (EAP). Los detalles de este procedimiento exigen conocimiento estadístico y se basan en que se toma los patrones de respuesta de cada encuesta y usando los parámetros de ítems “calibrados” con la población de Calibración se estima únicamente la variable latente del modelo estadístico considerando la distribución de probabilidad del modelo de respuesta graduada. Esta variable latente es la correspondiente medida TRI de la escala y es un puntaje alternativo al puntaje bruto. Este puntaje TRI está altamente correlacionado con el puntaje bruto y por tanto, estudiantes con medidas TRI altas tienen medidas brutas altas también y viceversa. Pero la medida TRI tiene mejores características que la medida bruta. Esta media está referenciada en una población de referencia en una escala de medida siguiendo la distribución normal con media 0 y desviación estándar 1

Afortunadamente el procedimiento esta implementado en muchos programas estadísticos libres como el programa R usado aquí (específicamente usamos el paquete mirtCAT) así como programas estadísticos como Stata, SAS y SPSS.

La medida obtenida bajo el modelo de respuesta graduada durante el proceso de estimación es una variable continua en una escala similar a un puntaje estandarizado tipo z basado en la

distribución normal estándar[23]. Para una presentación conveniente en valores positivos y su uso para análisis de tipo cuantitativo se puede transformar esta escala de medida inicial de puntaje z en una nueva escala de medida usando determinadas transformaciones. La escala transformada usada aquí es la escala de puntaje T (que tiene un centro o promedio ideal de 50 y una desviación estándar de 10)[24]. Considerando esta escala esperamos que las medidas obtenidas se concentren en su mayor parte entre 30 y 70 puntos[25].

Con la metodología adoptada, tendremos medidas en una escala que resulta comparable a otras medidas en el tiempo y en entre observaciones, la cual resulta independientemente del número de preguntas incluidas en cada medida y del tipo de medida considerada. Adicionalmente las desviaciones son interpretables como desvíos de la media de referencia adoptada.

Cabe aclarar que, en el caso de querer realizar inferencias para los diferentes grupos de estudiantes en una determinada población de interés, en otras palabras, para caracterizar una determinada población y realizar un análisis diagnóstico a nivel de grupos de interés se requiere obtener una representatividad de tipo estadístico. Por ejemplo, realizar comparaciones por sexo o por niveles de riesgo distrital requiere tener muestras representativas de cada grupo a ser comparado los cuales son los requerimientos usuales en la inferencia estadística, pero lo importante ahora es que contamos con medidas TRI continuas y entonces métodos estadísticos inferenciales suponiendo normalidad puede ser usados, incluyendo modelos de regresión, modelos de factores asociados y modelos de ecuaciones estructurales.

1.4. Definición de criterios de clasificación para niveles de desarrollo.

Uno de los objetivos del uso del Instrumento es poder describir a la población de estudiantes en términos del nivel en el que se encuentra (en promedio) en cuanto al bienestar socioemocional, de modo que la información pueda ser usada para interpretar los resultados a nivel de colegio y de localidad geográfica. Si bien, para cada estudiante se obtiene el correspondiente nivel

[23] Es decir, el resultado de este proceso nos permite obtener medidas en casi la totalidad de las observaciones en un rango de valores aproximados entre -3 y 3 centrado con un promedio 0 que corresponde al centro ideal de la medida del desempeño y con una desviación estándar de 1 punto.

[24] El puntaje T es obtenido usando la expresión T=50+10.Z, donde Z es el puntaje estandarizado. Este proceso es equivalente al usado por ejemplo por Laboratorio Latino Americano de Evaluación de la Calidad Educativa que usa un promedio 250 y una desviación estándar 50 (ver por ejemplo http://www.grade.org.pe/download/pubs/InvPolitDesarr-10.pdf) y similar a la transformación realizada por la Unidad de Medición de Calidad Educativa del Perú cuando por ejemplo reportan el desempeño de las Evaluación Censal 2013 usando promedio 500 y desviación estándar 100. Ver ttp://umc.minedu.gob.pe/wp-content/uploads/2014/03/ppt-resultados-web-UMC-ECE2013.-3demarzo-ult21.pptx

[25] Esto significa que a nivel teórico y no en una determinada muestra, esperamos que las medidas de las personas en la población obtenidas considerando el modelo de respuesta graduada estén en el intervalo de 30 a 70 puntos con un 95 % de probabilidad.

alcanzado, los resultados del Instrumento, sin embargo, no deben ser utilizados para hacer diagnósticos o interpretaciones a nivel individual.

Como es señalado por Crocker y Algina (2006), en muchos casos, es muy difícil realizar interpretaciones útiles o hacer inferencias de los puntajes de los evaluados solamente. Por lo tanto, en esta última parte del estudio se definirán puntos de corte que le permitirá a los usuarios del Instrumento entender mejor los resultados al poder determinar el porcentaje de la población que se encuentra en diferentes niveles de desarrollo socioemocional (por ejemplo, a nivel colegio, localidad, municipio o por distinguiendo por género, nivel socioeconómico, urbano/rural, etc.).

Dado que las medidas en el Instrumento por causa de la aplicación del modelo de respuesta graduada son asumidas normalmente en la población de referencia y dado que las medidas estimadas para cada evaluado son del tipo z o calificación estándar (puntaje t en nuestro caso) sin que sea necesario realizar un proceso de transformación para normalizarlas resulta más conveniente aprovechar esta característica para hacer uso directo de las calificaciones estándar[26] en la elaboración de los criterios de clasificación. Las calificaciones estándar expresan la distancia del individuo con respecto a la media en términos de la deviación estándar de la distribución normal.

La muestra de referencia o grupo normativo es la principal característica de un modelo de normas porque las calificaciones se realizan en relación a este grupo. Existen tres formas de definir la muestra normativa: a) Obtener una muestra con representatividad estadística, b) Considerar la población ideal obtenida por el modelo de respuesta graduada como muestra normativa, y c) definir una muestra normativa de referencia basada en otro estudio. El primer caso es el procedimiento usual dentro de un esquema de normas y requiere una muestra realmente representativa que en este momento no se dispone. El tercer caso supone la existencia de un estudio previo que pueda ser usado considerando medidas equivalentes o similares. Lamentablemente, tampoco es el caso aquí. En este estudio se ha decidido por seguir el caso b).

La propuesta b) para generar los criterios de clasificación, corresponde a una postura ecléctica y ya ha sido adoptada en un estudio similar en Perú (Bazán, 2014). A pesar de no contar, en un principio con una muestra representativa, a través del uso de pesos y remuestreo se ha procurado por obtener una población muy similar a la población de referencia (población escolar del país). Así, contamos con una población de referencia con medidas TRI que vienen de una distribución normal plenamente justificada[27].

[26] "Los instrumentos actuales hacen un uso creciente de las calificaciones estándar, que desde cualquier punto de vista constituyen el tipo más satisfactorio de puntuación derivada" (Anastasi y Urnina, 1998, pág 61.)

[27] Justificadas por el conjunto de análisis realizados en el primer estudio con respecto a la unidimensionalidad y confiabilidad de las escalas, y además, por los resultados del ajuste de los modelos TRI en el proceso de calibración que se desarrolló en la sección anterior.

Por tanto, los valores que se obtienen en el proceso de estimación de medidas toman como referencia la población ideal y por lo tanto esta población se constituye en un grupo normativo.

Por ejemplo, un valor de 40 puntos (μ-σ, donde μ=50 corresponde al promedio normativo de la población ideal y σ=10 corresponde a la desviación estándar de la población ideal) significa que la medida está por debajo del promedio de la población ideal y exactamente a una desviación estándar de dicha media.

Otra ventaja evidente del hecho que el modelo de respuesta graduada produzca medidas normales es la comparabilidad entre las medidas[28] porque para dos medidas diferentes, por ejemplo 40 puntos en cada una tienen el mismo significado: ambas están por debajo del promedio de la población ideal y exactamente a una desviación estándar de dicha media

En el siguiente cuadro se presenta la propuesta de norma o baremo para las medidas considerando como referencia la población ideal.

Criterio en la población ideal	Medidas	% esperado en la población ideal	Nivel
Medidas menores que μ-0.5σ	Medidas menores que 45	30.85	Bajo
Medidas entre μ-0.5σ y μ+0.5σ	Medidas entre 45 y 55 puntos	38.30	Medio
Medidas mayores que μ+0.5σ	Medidas mayores de 55	30.85	Alto

La propuesta se basa en una comparación de las medidas en relación al promedio del grupo normativo ideal o promedio esperado de la medida. Un puntaje menor de 45 es clasificado como bajo y significa que el evaluado obtuvo un nivel que corresponde a un puntaje menor que media desviación estándar por debajo del promedio esperado de la medida o su desempeño es equivalente al 30.85 % más bajo de las medidas en la población de referencia. De similar modo, un puntaje mayor a 55 es clasificado como alto y significa que el evaluado obtuvo un nivel que corresponde a un puntaje mayor que media desviación estándar por encima de del promedio esperado de la medida o su desempeño es equivalente al 30.85 % más alto de las medidas en la población de referencia.

Considerando esta propuesta, el grupo con bajos resultados puede ser denominado *grupo en riesgo* o de *bienestar bajo*, y el grupo alto puede ser denominado *grupo prosperando* o de

[28] "Las puntuaciones estándares derivadas linealmente sólo son comparables cuando provienen de distribuciones de más o menos la misma forma; por ejemplo, en tales condiciones, una calificación que corresponde a una desviación estándar por encima de la medida significa que el individuo ocupa la misma posición en los dos grupos. En ambas distribuciones, la calificación supera aproximadamente el mismo porcentaje de sujetos, y ese porcentaje puede ser determinado si se conoce la forma de la distribución" (Anastasi y Urbina, 1998), pág 62.

bienestar alto. En este caso, el grupo intermedio puede ser denominado *grupo en proceso* o con habilidades *insuficientes*, y así queda más claro que no es lo esperado sino que es un grupo en camino para el grupo de bienestar alto pues lo que se quiere en una población ideal es que la mayoría esté en el nivel alto (prosperando) y muy pocos en el nivel bajo. Las variaciones posibles en los puntajes de corte con este procedimiento antes que una desventaja constituye una herramienta poderosa para decisiones de política ya que son posibles la definición de nuevos estándares dependiendo de las decisiones de clasificación y metas que puedan ser definidos.

Estos puntajes de corte pueden ser modificados futuramente dependiendo de nuevas metas que se puedan proponer, por ejemplo, estableciendo puntos de corte más altos para discriminar el *grupo prosperando* conforme se observen mejoras entre los estudiantes como consecuencia de decisiones de política, intervenciones oportunas y trabajo de las escuelas. En una situación ideal se espera que la mayoría de estudiantes de las instituciones educativa alcance un adecuado desarrollo Socio-Emocional y recursos extra escolares, se observe un óptimo clima escolar y disminución de la Violencia escolar.

Capítulo 3. Conclusiones y sugerencias

Conclusiones

1. Con el fin de mejorar la representatividad y la precisión de las estimaciones de parámetros y medidas se ha optó por utilizar, por un lado, pesos de tal forma que la muestra efectiva se asimile lo más posible a la población escolar de Colombia, y por otro lado, técnicas de remuestreo para garantizar la robustez de los resultados de futuras aplicaciones del Instrumento.

2. Las escalas del Instrumento en sus tres versiones fueron ajustadas al modelo de respuesta graduada. Los resultados del estudio de calibración de cada una de las preguntas del Instrumento, presentados en el Anexo 4, muestran que este modelo fue ajustado satisfactoriamente para cada una de estas escalas.

3. Considerando el proceso de calibración se dispone de una metodología para estimación de las medidas de los evaluados. Esta metodología substituye a la tradición calificación de las preguntas basada en la suma de las respuestas por un proceso computacional en el que son tomados en cuenta las diferencias entre las preguntas y un determinado modelo probabilístico. El resultado del proceso es un puntaje tipo Z. Este puntaje finalmente se transforma en una medida que presenta promedio 50 y desviación estándar 10.

4. La medida obtenida puede ser usada para cualquier estudio cuantitativo en reemplazo de los tradicionales puntajes suma. Esta medida es continua e idealmente proviene de una distribución normal. Con estas condiciones, esta medida es claramente una medida útil para cualquier proceso de análisis estadístico avanzado. Adicionalmente todas las medidas son comparables porque todas usan la misma escala de valores. Esto quiere decir, por ejemplo, que una diferencia de 5 puntos en la escala 1 es equivalente a la misma diferencia en cualquier otra escala.

5. Para propósitos de Clasificación se optó por una postura ecléctica asumiendo la formación de baremos basados en una muestra de referencia. El procedimiento de clasificación emplea la población de referencia ideal de las medidas que es normal y entonces definiendo desvíos en relación al promedio esperado basados en la desviación estándar se determine puntos de corte y porcentajes ideales en los grupos de clasificación.

6. La propuesta de clasificación presenta resultados comparables, indicando que un evaluado se encuentra en un nivel bajo cuando tiene una medida con valor menor a 45, y por otro lado se encuentra en un nivel alto cuando tienen una medida con valor mayor a 55.

7. Considerando el procedimiento para clasificar sugerimos crear tres grupos: grupo en riesgo (medidas menores que una desviación estándar del promedio ideal o puntaje menor que 40), grupo en proceso (medidas entre 40 y 50 puntos) y grupo prosperando (medidas encima del promedio ideal o mayor de 50). Estos puntajes de corte pueden ser modificados futuramente dependiendo de nuevas metas que se puedan proponer.

Sugerencias

1. Realizar estudios con una muestra representativa de las poblaciones escolares, que tengan en cuenta la diversidad de contextos de las instituciones educativas públicas y privadas para mostrar evidencias adicionales de la robustez del estudio presentado.

2. Realizar estudios de validación con criterios externos pueden ser recomendados futuramente para analizar la pertinencia de la clasificación propuesta en este estudio.

3. La metodología presentada es robusta por el proceso riguroso llevado a cabo sin embargo evidencia adicional acerca de la validez predictiva de los resultados puede ser considerada en futuros estudios.

Referencias

- AERA, APA, NCME (1999). Standars for educational and psychological testing. Preparado por un Comité conjunto de la American Educational Research Association, American Psychological Association, y el National Council on Measurement in Education. Washington: AERA.
- Anastasi, A. y Urbina, S. (1998) Test. Psicolgicos. (7ma Edición). Mxico: Prentice Hall. Department of Psycholpgy, Fordham University.
- Baker, F. B., Kim, S.-H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.
- Bazán J. (2014). Análisis psicométrico del Cuestionario de Bienestar Escolar de Primaria y Secundaria. Estudio 2014. Informe elaborado por el Dr. Jorge Luis Bazán por encargo del Banco Mundial-Perú, bajo la supervisión de Inés Kudó (Gerente del Programa de Asistencia Técnica en educación).
- Bazán, J. L (2011). Análisis psicométrico de EGRA y su validez concurrente con otras evaluaciones de desempeño en lectura: caso Honduras y Nicaragua. Ed Data II". Informe preparado para la Oficina de Desarrollo Económico, Agricultura y Comercio (EGAT/ED) y para la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID). Disponible en <https://www.eddataglobal.org/data/index.cfm/ValidezConcurrente_final_15marzo2011.pdf?fuseaction=throwpub&ID=301>
- Bazán, J. y Millones, J. (2002) Evaluación psicométrica de las preguntas de las escalas CRECER 98. En Rodriguez, J. , Vargas, S. (eds). Análisis de los Resultados y Metodología de las Escalas Crecer 1998. Documento de trabajo 13. Lima: MECEP-Ministerio de Educación. Pp: 141-170. Disponible en http://www.minedu.gob.pe/umc/publicaciones/mecep/doc13/13h.pdf
- Bazán J., Merino, M. H. e Mazzon, J. A. (2011). Classificação de modelos de resposta ao item policotômicos com aplicação ao marketing. Revista Brasileira de Estatística, 72, 7-39.
- Bazán J. (2014). Análisis psicométrico del Cuestionario de Bienestar Escolar de Primaria y Secundaria. Documento no publicado.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Statistical Theories of Mental Test Scores, ed. F. M. Lord and M. R. Novick, 395–479. Reading, MA: Addison–Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443-459.
- Burga, A. (2005). La unidimensionalidad de un instrumento de medición: perspectiva factorial. Disponible en < http://www2.minedu.gob.pe/umc/admin/images/publicaciones/artiumc/2.pdf>
- Carmines, Edward G. y Richard A. Zeller. (1979). Reliability and Validity Assessment. Beverly Hills, CA: Sage.

- Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi: 10.18637/jss.v048.i06
- Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25, 31-45
- Crocker, L., Algina, J. (2006). Introduction to classical and modern test theory. Belmont, CA: Wadsworth Publishing Company.
- Gadermann, A.M., Guhn, M., y Zumbo, B.D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. Practical Assessment, Research, & Evaluation, 17 (3), 1-13.
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
- Joint Committee on Standards for Educational Evaluation. (2003). The Student Evaluation Standards: How to Improve Evaluations of Students. Newbury Park, CA: Corwin Press. Disponible en < http://www.wmich.edu/evalctr/jc/briefing/ses/>
- Knol DL, Berger MP. (1991). Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26, 457-477.
- Nering, M. Ostini, R. (2011). Handbook of Polytomous Item Response Theory Models. Abingdon, Oxon: Routledge.
- Lord, F.M., Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.
- Orlando, M., Sherbourne, C.D., Thissen, D. (2000) Summed-score linking using item response theory: Application to depression measurement, Psychological Assessment, 12(3), 354-359.
- Revelle, W. (2012). Psych. Procedures for personality and psychological research. R package version 1.28.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, **34**, 100--114.
- Schonert-Reichl, Kimberly A.; Guhn, Martin; Gadermann, Anne M.; Hymel, Shelley; Sweiss, Lina; Hertzman, Clyde (2012). Development and Validation of the Middle Years Development Instrument (MDI): Assessing Children's Well-Being and Assets across Multiple Contexts. Social Indicators Research : 1-25
- Tay L, Meade AW, Cao M. (2015). An overview and practical guide to IRT measurement equivalence analysis. Organ Res Meth.. doi:10.1177/ 1094428114553062.
- Wilson, M. (2005). Constructing Measures: An Item Response Modeling Approach. Lawerence Erlbaum Associates, Inc Publisher.

- Zickar, M., Russell, S., Smith, C., Bohle, P. y Tilley, A. (2002). Evaluating two morningness scales with item response theory. Personality and Individual Differences, 33, 11-24.
- Zumbo, B. D., Gadermann,A. M., y Zeisser, C.. (2007). Ordinal Versions of Coefficients Alfa and Theta For Likert Rating Scales. Journal of Modern Applied Statistical Methods, 6, 21-29.

Anexo 1. Evidencias de Validez de los Cuestionarios

Tabla 8: Fuentes de las escalas

Nombre escala	Instrumento original	Fuente original de la escala
Acoso escolar	Aulas en Paz	Velásquez, A. M., & Chaux, E. (2005). AVC–agresión, violencia & competencias. Bogotá: Universidad de los Andes. Unpublished document.
Agresión escolar	CuBE	Psychometric Support for an Abbreviated Version of the California School Climate and Safety Survey Jennica L. Rebelez and Michael J. Furlong In press (a revised version of this article is being type set at this time). International Journal of School and Educational Psychology
Asertividad	Aulas en Paz	Velásquez, A.M. (2005). Desarrollo de la asertividad: Comparación entre dos intervenciones pedagógicas. Master's Thesis in Education. Bogotá: Universidad de los Andes.
Autoeficacia académica	CuBE	Roeser, W.R., Midgley, C., & Urdan, T.C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: the mediating role of goals and belonging. Journal of Educational Psychology, 88(3), 408- 422.
Autopercepción general	CuBE	Marsh, H. W. (1988). Self- Description Questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self- concept: A test manual and a research monograph. San Antonio, Texas: The Psychological Corporation.
Conducta prosocial	CuBE	Youth Outcome Measures for AfterSchool KidzLit™, Developmental Studies Center, 2001
Diversidad (acciones)	Adaptación ICFES	Banco de ítems ICFES
Diversidad (actitudes)	Adaptación ICFES	Banco de ítems ICFES
Empatía	Aulas en Paz	Schulz, W., Ainley, J., Friedman, T., & Lietz, P. (2011). ICCS 2009 Latin American Report. Civic knowledge and attitudes among lower- secondary students in six Latin American countries. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA). ACER. Nfer. Università degli Studi Roma Tre.
Género	Adaptación ICFES	Banco de ítems ICFES
Grit	Adaptación Grit	Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). Journal of personality assessment, 91(2), 166-174.
Intimidad en la amistad	CuBE	Hayden-Thomson, L. K. (1989). The development of the Relational Provisions Loneliness Questionnaire for children. Unpublished doctoral dissertation, University of Waterloo, Waterloo, Ontario, Canada.
Proyecto de vida	Fundación Corona	Fundación Corona (2018). Programa de Orientación Socio Ocupacional, Guía De Evaluación. Sin publicar.
Regulación emocional	Aulas en Paz	Velásquez, A. M., & Chaux, E. (2005). AVC–agresión, violencia & competencias. Bogotá: Universidad de los Andes. Unpublished document.

Seguimiento de la ley	Adaptación ICFES	Banco de ítems ICFES
Sentido de pertenencia con pares	CuBE	Hayden-Thomson, L. K. (1989). The development of the Relational Provisions Loneliness Questionnaire for children. Unpublished doctoral dissertation, University of Waterloo, Waterloo, Ontario, Canada.
Toma de decisiones responsables	Adaptación ICFES	Banco de ítems ICFES
Trabajo en equipo	Adaptación ICFES	Banco de ítems ICFES

Tabla 9: Evidencias de confiabilidad y unidimensionalidad del instrumento para primaria basadas en el estudio piloto 2019

Escala	Preguntas	Análisis de confiablidad	Análisis de unidimensionalidad		
		Alfa ordinal	RSMR	Varianza explicada 1er factor	Correlación de scores con factores
Acoso escolar	5	0.84	0.09	0.51	0.91
Agresión escolar	5	0.84	0.02	0.52	0.92
Asertividad	4	0.74	0.01	0.43	0.86
Autoeficacia académica	3	0.78	0.00	0.54	0.84
Conducta prosocial	3	0.72	0.00	0.46	0.85
Diversidad (acciones)	6	0.90	0.06	0.59	0.95
Diversidad (actitudes)	7	0.89	0.05	0.53	0.94
Empatía	7	0.94	0.02	0.68	0.97
Género	7	0.89	0.05	0.47	0.95
Grit	3	0.70	0.00	0.44	0.84
Intimidad en la amistad	3	0.78	0.00	0.55	0.89
Regulación emocional	4	0.77	0.03	0.47	0.88
Sentido de pertenencia con pares	3	0.72	0.00	0.46	0.85
Toma de decisiones responsables	4	0.79	0.05	0.49	0.89
Trabajo en equipo	6	0.83	0.05	0.38	0.91

Nota i) Los valores del coeficiente Alfa son aceptables si son superiores a 0.7, buenos para valores superiores a 0.8 y excelentes para valores superiores a 0.9; ii) Para que una escala sea considerada unidimensional se deben cumplir al menos 2 de los 3 criterios señalados en las últimas 3 columnas de la tabla; correlación de escores con factores mayores que 0.85, 40 % de proporción de variancia de las preguntas es explicada por el primer factor y la raíz cuadrada del error cuadrático medio ser menor que 0.05.

Tabla 10: Evidencias de confiabilidad y unidimensionalidad del instrumento para secundaria basadas en el estudio piloto 2019

Escala	Preguntas	Análisis de confiablidad	Análisis de unidimensionalidad		
		Alfa ordinal	RSMR	Varianza explicada 1er factor	Correlación de scores con factores
Acoso escolar	5	0.89	0.06	0.62	0.95

Escala	Preguntas	Análisis de confiablidad	Análisis de unidimensionalidad		
		Alfa ordinal	RSMR	Varianza explicada 1er factor	Correlación de scores con factores
Agresión escolar	6	0.85	0.03	0.49	0.92
Asertividad	4	0.80	0.01	0.51	0.90
Autoeficacia académica	3	0.75	0.00	0.50	0.87
Autopercepción general	3	0.72	0.00	0.47	0.85
Conducta prosocial	3	0.75	0.00	0.50	0.86
Diversidad (acciones)	8	0.92	0.04	0.58	0.96
Diversidad (actitudes)	7	0.90	0.04	0.54	0.95
Empatía	7	0.93	0.04	0.66	0.97
Género	7	0.94	0.04	0.61	0.97
Grit	3	0.78	0.00	0.54	0.88
Intimidad en la amistad	3	0.86	0.00	0.68	0.93
Regulación emocional	6	0.80	0.04	0.41	0.90
Sentido de pertenencia con pares	3	0.76	0.00	0.52	0.87
Toma de decisiones responsables	4	0.84	0.07	0.57	0.92
Trabajo en equipo	6	0.86	0.05	0.46	0.93

Nota i) Los valores del coeficiente Alfa son aceptables si son superiores a 0.7, buenos para valores superiores a 0.8 y excelentes para valores superiores a 0.9; ii) Para que una escala sea considerada unidimensional se deben cumplir al menos 2 de los 3 criterios señalados en las últimas 3 columnas de la tabla; correlación de escores con factores mayores que 0.85, 40 % de proporción de variancia de las preguntas es explicada por el primer factor y la raíz cuadrada del error cuadrático medio ser menor que 0.05.

Tabla 11: Evidencias de confiabilidad y unidimensionalidad del instrumento para media basadas en el estudio piloto 2019

Escala	Preguntas	Análisis de confiablidad	Análisis de unidimensionalidad		
		Alfa ordinal	RSMR	Varianza explicada 1er factor	Correlación de scores con factores
Acoso escolar	5	0.92	0.05	0.71	0.96
Agresión escolar	6	0.87	0.04	0.53	0.93
Asertividad	4	0.79	0.02	0.50	0.89
Autoeficacia académica	5	0.85	0.03	0.54	0.92
Autopercepción general	3	0.79	0.00	0.55	0.89
Conducta prosocial	3	0.76	0.00	0.52	0.87
Diversidad (acciones)	8	0.92	0.06	0.59	0.96
Diversidad (actitudes)	7	0.92	0.04	0.64	0.96
Empatía	7	0.93	0.05	0.67	0.97
Género	7	0.95	0.04	0.67	0.97
Grit	3	0.82	0.00	0.62	0.91
Intimidad en la amistad	3	0.87	0.00	0.71	0.94
Proyecto de vida	5	0.84	0.09	0.51	0.92

Regulación emocional	7	0.79	0.05	0.37	0.89
Seguimiento de la ley	4	0.85	0.08	0.59	0.92
Sentido de pertenencia con pares	3	0.81	0.00	0.59	0.90
Toma de decisiones responsables	4	0.87	0.06	0.64	0.94
Trabajo en equipo	6	0.87	0.06	0.49	0.93

Nota i) Los valores del coeficiente Alfa son aceptables si son superiores a 0.7, buenos para valores superiores a 0.8 y excelentes para valores superiores a 0.9; ii) Para que una escala sea considerada unidimensional se deben cumplir al menos 2 de los 3 criterios señalados en las últimas 3 columnas de la tabla; correlación de escores con factores mayores que 0.85, 40 % de proporción de variancia de las preguntas es explicada por el primer factor y la raíz cuadrada del error cuadrático medio ser menor que 0.05.

Anexo 2. Características psicométricas de los Cuestionarios

1. Validez de los cuestionarios

Como es dicho en Bazán (2011), la validez se refiere al grado por el cual la evidencia y la teoría respaldan las interpretaciones de los puntajes de los cuestionarios. La validez es el aspecto más importante en el desarrollo y evaluación psicométrica de estas.

Es por eso que el proceso de validación involucra la acumulación de evidencia para proveer una base científica para las propuestas de interpretación de los puntajes de un cuestionario. De esta manera se deben evaluar los usos propuestos del cuestionario.

Cómo se ha dicho antes, la decisión acerca de qué tipo de evidencias son importantes para validación varían por cada tipo de situación de evaluación. Sin embargo, es natural que algunos tipos de evidencia puedan ser especialmente críticos en un determinado caso, mientras que otros pueden ser menos útiles.

La validación implica poner gran atención sobre las posibles distorsiones en el significado alrededor de una representación inadecuada del constructo, así como en los aspectos de medición tales como el formato, las condiciones de administración o el nivel de lenguaje, todos ellos aspectos que pueden limitar o modificar la interpretación de los puntajes de la prueba.

El proceso de validación puede llevar a revisiones del cuestionario, su marco conceptual o ambos.

La validación es una responsabilidad conjunta del desarrollador y del usuario del cuestionario. El desarrollador es responsable de dar evidencia relevante y una racionalidad que respalde el uso pretendido del cuestionario. Por otro lado, el usuario es el responsable final de evaluar la evidencia con respecto al particular propósito de uso del cuestionario. Cuando el uso de un cuestionario difiere de lo respaldado por el desarrollador, el usuario es especialmente responsable de la validación.

Hay diferentes tipos de evidencia de validez, que se pueden clasificar en:

- Evidencias basadas en el contenido de la prueba.
- Evidencias basadas en el proceso de respuesta.
- Evidencias basadas en la estructura interna.
- Evidencias basadas en las relaciones con otras variables.
- Evidencias basadas en las consecuencias del uso de la prueba.

a. *Evidencias basadas en el contenido de la prueba:* Se obtienen a partir de un análisis de la relación entre el contenido del cuestionario y el constructo a ser medido. El contenido se refiere a temas, redacciones, formatos de preguntas, tareas o cuestiones en los cuestionarios, así como a las guías para los procedimientos con respecto a administración y calificación.
   La evidencia puede incluir un análisis lógico y empírico de la adecuación, con la cual el contenido del cuestionario representa el dominio de contenido, y de la relevancia del dominio de contenido para la interpretación de los puntajes. También puede incluir juicio de expertos de la relación entre las partes del cuestionario y el constructo.
b. *Evidencias basadas en los procesos de respuesta:* Análisis teóricos y empíricos de los procesos de respuesta proporcionan evidencias referidas al ajuste entre el constructo y la naturaleza del desempeño de respuesta que usan los examinados..
   Este tipo de evidencia generalmente proviene de análisis individuales de respuestas, o de cuestiones referidas a las estrategias de resolución de las preguntas, incluyendo tiempos de respuesta. También se puede obtener evidencia analizando la relación entre partes del cuestionario y otras variables que ayudan a reconsiderar formatos.
   Estudios de procesos de respuesta que involucran a examinados de diferentes subgrupos, pueden ayudar a determinar la influencia de otras variables en el desempeño durante la aplicación del cuestionario.
c. *Evidencias basadas en la estructura interna del cuestionario:* Estas evidencias pueden indicar el grado en el que la relación entre las preguntas del cuestionario y sus componentes conforman el constructo en el cual se basan los puntajes que surgen del cuestionario.
   El marco conceptual de un cuestionario puede implicar una simple dimensión o varios componentes que se esperan homogéneos pero distintos entre sí.
d. *Evidencias basadas en relaciones con otras variables:* El análisis de los puntajes del cuestionario con otras variables externas proporciona una fuente importante de validez. Variables externas pueden ser: medidas del mismo criterio que el cuestionario espera predecir, otros cuestionarios que hipotéticamente miden el mismo constructo, cuestionarios que miden constructos relacionados o diferentes.
e. *Evidencias basadas en las consecuencias de la prueba:* En años recientes se ha prestado atención a la incorporación de las consecuencias pensadas y no pensadas del cuestionario en el aspecto de validez. Aunque la información acerca de las consecuencias del cuestionario pueda influir en las decisiones de su uso, tales consecuencias no deben disminuir la validez de las interpretaciones planificadas. Los cuestionarios se administran, comúnmente, en la expectativa de obtener algún beneficio con sus puntajes. Así, un propósito fundamental de la validación es indicar cómo serán obtenidos estos beneficios.

Un argumento de validez legítimo debe integrar varias fuentes de evidencia en una cantidad coherente con el grado en el cual la evidencia existente y la teoría respaldan las interpretaciones pretendidas de los puntajes del cuestionario, y abarcar evidencia recogida de nuevos estudios y aquella disponible de reportes de investigación recientes.

El argumento de validez puede indicar la necesidad de refinar la definición del constructo, puede sugerir revisiones en la prueba u otros aspectos del proceso de pruebas y puede indicar áreas que necesitan estudios futuros.

Finalmente, la validez de una interpretación intencional de los puntajes de los cuestionarios cuenta con toda la evidencia relevante disponible para la calidad técnica del sistema. Esto incluye evidencia de aspectos relativos a la construcción de la prueba, como:

- confiabilidad adecuada
- administración de la prueba apropiada,
- calificación de la prueba apropiada,
- escalamiento de puntajes precisos,
- equivalencias,
- protocolos estandarizados,
- atención cuidadosa respecto a la honestidad de las respuestas de los examinados.

Para asegurar estos aspectos, AERA, APA, NCME (1999) listan 24 estándares de validez.

2. Análisis satisfactorio de las preguntas de los cuestionarios

En general, el objetivo en el análisis de las preguntas es eliminar las preguntas que no satisfacen un conjunto de propiedades que hacen de la prueba un instrumento eficaz y sin sesgo en la estimación de la dimensión medida. Estas propiedades están supeditadas al marco referencial teórico (modelo) que se adopte para el análisis de los cuestionarios. El principal análisis de preguntas basado en el enfoque clásico considera el análisis de la consistencia interna que incluye el reporte de la correlación ítem-total y el alfa de Cronbach si el ítem se elimina.

Sin embargo existen otros tipos de análisis como a) el análisis de datos perdidos[29], b) análisis de patrones de respuesta[30] y c) el análisis de la homogeneidad y discriminación[31] que son usados especialmente cuando se desarrolla un cuestionario en sus etapas iniciales.

3. Confiabilidad de cuestionarios

Un cuestionario, definida de manera amplia, es un conjunto de tareas o una escala, diseñadas para describir o hacer explícitas conductas de examinados en un dominio específico, o un

[29] e.g., altos grados de datos perdidos en las preguntas pueden sugerir que una pregunta es difícil, no clara o incomprensible.

[30] e.g., chequeando únicamente las altas o bajas opciones de respuesta puede reflejar una pérdida de comprensión o interés); una medida apropiada en este caso es reporta la media y el coeficiente de variabilidad de las preguntas (desviación estándar/media).

[31] Análisis de la homogeneidad y discriminación incluye el análisis de la asimetría de las respuestas (para examinar si los evaluados usan la escala entera y/o mirar que ítems discriminan entre estudiantes.

sistema para recolectar muestras de trabajos individuales en un área particular. Acoplado a este dispositivo hay un procedimiento de calificación que hace posible que el examinador pueda cuantificar, evaluar, e interpretar las muestras de conducta o trabajo.

La confiabilidad se refiere a la consistencia de tales medidas cuando los procedimientos de *evaluaciones* son repetidos en poblaciones de individuos o grupos admitiendo la presencia de un componente de error.

Decir que un puntaje implica un componente de error significa que existe un hipotético valor libre de error que caracteriza a un examinado al momento de la evaluación. Por ejemplo, en la Teoría Clásica de los Test este valor es el *puntaje verdadero* (puntaje promedio hipotético resultante de muchas repeticiones de la prueba o formas alternativas del instrumento).

La diferencia hipotética entre el puntaje observado del examinado en cualquier medición particular y el puntaje verdadero o universal es llamada “error de medición”.

Nuevamente, para asegurar estos aspectos, AERA, APA, NCME (1999) listan 20 estándares acerca de confiabilidad y error de medición.

De acuerdo con AERA, APA, NCME (1999), la confiabilidad de un cuestionario mide el grado en que el cuestionario es consistente en los puntajes que de ella se obtienen. Idealmente se determina tomando dos o más veces el mismo cuestionario a un examinado y revisando si los puntajes obtenidos son consistentes (idénticos o similares). En la práctica, la consistencia se determina de formas alternativas, una de las cuales se basa en la consistencia interna del cuestionario; por ejemplo, cuán consistentemente mide la mitad de una prueba respecto a su otra mitad. Este criterio de consistencia interna puede ser calculado por el coeficiente “Alfa” de Cronbach y es reportado comúnmente en diversos estudios psicométricos.

El coeficiente Alfa de Cronbach es un índice que da un valor o cota inferior a la verdadera confiabilidad, pero no es el único. Adicionalmente puede establecerse que: 1) en la práctica el Alfa de Cronbach proporciona valores que están fuera del rango de valores de confiabilidad derivados de una sola administración, 2) Alfa es más usado por estar disponible en software comerciales y por ser reconocido en la comunidad académica, en detrimento de otros índices propuestos. Recientemente Gadermann, Ghun y Zumbo (2012) y Zumbo, Gadermann y Zeisser, (2007) han propuesto el cálculo del coeficiente alfa de Cronbach ordinal que mejora sustantivamente la precisión en la estimación del alfa considerando la matriz de correlaciones policóricas en vez de la matriz de correlaciones de Pearson.

4. Unidimensionalidad de los cuestionarios

El propósito principal de los cuestionarios es estimar las diferentes áreas que lo componen y en este caso suponemos que el constructo correspondiente al área es una variable latente subyacente al conjunto de respuestas dados a las preguntas de esta área.

En el esquema moderno del concepto de validez se incluye la evidencia de unicidad, es decir, la propiedad de un factor o área de un cuestionario medir únicamente un constructo o desempeño (unicidad de la prueba medible), esto es, establecer si el conjunto de preguntas dentro de cuestionario mide una sola cosa -es decir evaluar la unidimensionalidad[32].

No siempre es posible determinar que este supuesto se cumpla cabalmente sin embargo se requiere evaluarlo. Algunas maneras de evaluar la unidimensionalidad surgen a partir de la matriz de correlaciones tetracóricas de las preguntas que componente una prueba considerando. Entre estas posibilidades figura un análisis factorial a partir de la matriz de correlaciones policóricas (Knol & Berger, 1991) y también recientemente el paquete psych del programa R (Revelle, 2012) implementa un análisis factorial exploratorio bajo la metodología de Teoría de Respuesta al Ítem que puede ser empleado también.

Dado que la unidimensionalidad es el supuesto más importante para justificar la medición del área, en general, para tomar una decisión acerca de la violación de este supuesto es importante considerar no uno, sino varios criterios para rechazarla[33]. Entre estos criterios generalmente usados podemos indicar el Análisis Factorial de Mínimos cuadrados no ponderados o de mínimos residuales con rotación oblimin usando matriz de correlaciones policóricas.

Según Carmines y Zeller (1979) puede considerarse que una prueba será unidimensional pese a presentar varios factores si el primer factor explica por lo menos el 40% de la varianza. También Orlando et al (2000), indican que un conjunto de preguntas puede tener múltiples valores propios superiores a 1 y por tanto más de un factor pero aun así puede ser lo suficiente unidimensional para ser analizado con un modelo de teoría de respuesta al ítem. Así dichos autores consideran que si el número de preguntas con cargas factoriales superior a 0,35 es bastante alto -digamos superior a 80 %- esto puede ser considerado como una evidencia aceptable de unidimensionalidad.

Complementariamente podemos considerar dos medidas adicionales, la raíz del cuadrado medio de residuales (root mean square of the residuals (RMSR) en inglés) que esperamos sea menor que 0.05 así como la correlación del puntaje con el factor que esperamos sobre el valor de 0.85. Finalmente una escala será considerada unidimensional si cumple dos de tres criterios, correlación de escores con factores mayores que 0.85, 40 % de proporción de variancia en el primer factor y RSME<0.05.

5. El uso de un modelo TRI

Los supuestos básicos que exige todo modelo TRI son básicamente cuatro:

[32] Una interesante discusión acerca de la unidimensionalidad y de las maneras como evaluarla puede verse también en Burga (2005) y en Abedi (1997).

[33] Algunos autores consideran una prueba como unidimensional cuando al menos un criterio se satisface y como multidimensional cuando varios criterios no se satisfacen simultáneamente.

1. **Monotonicidad:** el supuesto indica que a medida que aumenta el nivel de la habilidad, también aumenta la probabilidad de una respuesta correcta.
2. **Unidimensionalidad:** el modelo asume que se está midiendo una variable latente dominante y que esta variable es la fuerza impulsora de las respuestas observadas para cada ítem de la medida.
3. **Independencia local:** que la respuesta a un ítem no influye en la respuesta dada a ningún otro. Esto permite afirmar que la probabilidad de responder correctamente a un conjunto de ítems es el producto de las probabilidades de contestar correctamente a cada pregunta por separado.
4. **Invarianza:** se nos permite estimar los parámetros de los ítems desde cualquier posición en la curva de respuesta al item. En consecuencia, podemos estimar los parámetros de un ítem a partir de cualquier grupo de sujetos que hayan respondido al ítem.

Anexo 3. Modelo de respuesta graduada y modelo logístico de dos parámetros

El modelo de respuesta graduada y el modelo de logístico de dos parámetros son modelos de Teoría de Respuesta al Ítem que asumen que las respuestas a un ítem binario (de dos categorías de respuesta en dónde una se considera correcta) o politómico (que tiene varias categorías de respuesta ordenadas) pueden ser modeladas probabilísticamente identificado dos fuentes de explicación para esas probabilidades. Una asociada con las características de los propios ítems (llamado parámetros de ítem) y otra con la propia medida que está queriendo ser determinada (llamada rasgo latente). Ambas actúan de modo que, dependiendo del ítem y la medida, las personas terminan escogiendo una determinada categoría de respuesta.

Por ejemplo, en la siguiente pregunta del Instrumento tenemos 5 categorías de respuesta ordenadas. Como la pregunta corresponde a una pregunta del constructo Autopercepción general, esperamos que dependiendo de la medida que la persona tiene en este constructo puede inclinarse por escoger una determinada categoría de respuesta.

	¡NO!	No	Más o menos	Sí	¡Sí!
1. En general, me gusta mi manera de ser.	☐$_0$	☐$_1$	☐$_2$	☐$_3$	☐$_4$

Sin embargo, dado que varias otras preguntas son presentadas acerca del mismo constructo, por ejemplo, la pregunta 2, esperamos que una persona conteste una determinada opción de respuesta también en función de las propias características de las preguntas (por sus parámetros de ítem).

	¡NO!	No	Más o menos	Sí	¡Sí!
2. En general, tengo mucho de qué sentirme orgulloso(a).	☐$_0$	☐$_1$	☐$_2$	☐$_3$	☐$_4$

Para analizar estas situaciones y mejorar la metodología tradicional que consiste en formar los puntajes fila (suma de las categorías de respuesta) sin tomar en cuenta la diferencia entre las preguntas, se utiliza el modelo logístico de dos parámetros de Birnbaum (Birnbaum, 1968) y el Modelo de respuesta graduada de Samejima (Samejima, 1969) que serán descritos a continuación.

Modelo logístico de 2 parámetros para ítems binarios

Sea Y_{ij} el resultado por observar para el item i de la persona j, y y_{ij} el valor observado de Y_{ij}. Sin perder generalidad, se usará el termino correcto e incorrecto en referencia al resultado de Y_{ij}. Además, se referirá a $y_{ij} = 1$ como correcto y $y_{ij} = 0$ como incorrecto.

Usando la parametrización de TRI, se ve que la probabilidad de la persona j con nivel de rasgo latente θ_j de proveer una respuesta correcta al item i está dada por

$$\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = \frac{\exp exp\left\{a_i(\theta_j - b_i)\right\}}{1+\exp exp\left\{a_i(\theta_j - b_i)\right\}} \qquad \theta_j \sim N(0,1)$$

En dónde a_i y b_i representa la discriminación y la dificultad del ítem i respectivamente. Las dificultades representan un punto en el que una persona con nivel de rasgo latente $\theta_j = b_i$ tiene una probabilidad de 50% de responder correctamente el ítem i. Las estimaciones del parámetro de dificultad corresponden al punto del continuo del rasgo latente en dónde $\Pr Pr\,(\theta) = 0.5$. Debido a que se asume una media igual a 0 para θ un ítem se considera relativamente fácil sí es negativo y relativamente difícil si es positivo. Por otro lado, el parámetro de discriminación a_i da una idea de que tan correlacionado está la probabilidad de acertar un ítem con el nivel del rasgo latente. Es decir que entre más alto sea el parámetro de discriminación de un ítem, mayor será la probabilidad de dar una respuesta correcta a ese ítem en cuanto mayor sea el nivel rasgo latente de una persona (y viceversa).

Usando la forma intercepto-pendiente, la probabilidad de proveer una respuesta correcta se parametriza como

$$\Pr Pr\left(\alpha_i,\ \beta_i,\ \theta_j\right) = \frac{\exp exp\left(\alpha_i\theta_j + \beta_i\right)}{1+\exp exp\left(\alpha_i\theta_j + \beta_i\right)}$$

La transformación de estás dos parametrizaciones es

$$a_i = \alpha_i \qquad b_i = -\frac{\beta_i}{\alpha_i}$$

Sea $P_{ij} = \Pr Pr\left(\alpha_i,\ \beta_i,\ \theta_j\right)$ and $q_{ij} = 1 - P_{ij}$. Condicional en θ_j, las respuestas a los ítems se asumen independientes de tal manera que la densidad condicional para la persona *j* está dada por

$$f\left(\mathrm{B},\ \theta_j\right) = \prod_{i=1}^{I} p_{ij}^{\ y_{ij}} q_{ij}^{\ 1-y_{ij}}$$

Dónde $y_j = \left(y_{1j}, \ldots, y_{Ij}\right)$, $\mathrm{B} = (\alpha_1,\ \ldots, \alpha_I, \beta_1, \ldots, \beta_I)$, e I se refiere al número de ítems. La verosimilitud para la persona j es calculada al integrar la variable latente de la densidad conjunta

$$L_j(\mathrm{B}) = \int_{-\infty}^{\infty} f\left(\mathrm{B},\ \theta_j\right)\phi(\theta_j)d\theta_j$$

En dónde $\phi(\bullet)$ es la función de densidad para la distribución normal estándar, el logaritmo de la función verosimilitud para la muestra es la suma de los logaritmos de las funciones de verosimilitud de las N personas en la muestra.

$$logL(\mathrm{B}) = \sum_{j=1}^{N} logL_j(\mathrm{B})$$

La integral en la fórmula para $L_j(\mathrm{B})$ es generalmente no trazable, por lo que se deben usar métodos numéricos.

Modelo de respuesta graduada para ítems politómicos ordinales[34]

Sea Y_{ij} el resultado por observar para el ítem i de la persona j, y y_{ij} el valor observado de Y_{ij}. Sin perder generalidad, se asume que todos los ítems son de categorías ordenadas, $k = 0, 1, \ldots, K$.

Usando la parametrización de TRI, se ve que la probabilidad de la persona j con nivel de rasgo latente θ_j de proveer una respuesta k o mejor al ítem i está dada por

$$\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = \frac{\exp exp\left\{a_i(\theta_j - b_{ik})\right\}}{1+\exp exp\left\{a_i(\theta_j - b_{ik})\right\}} \qquad \theta_j \sim N(0,1)$$

En dónde a_i representa la discriminación del ítem i, $b_i = (b_{i1}, \ldots, b_{iK})$ representa las dificultades que distinguen las categorías ordenadas del ítem i, y se entiende que $\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = 1$ y $\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = 0$. La probabilidad de observar el resultado k es entonces

$$\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = \Pr Pr\left(a_i,\ b_i,\ \theta_j\right) - \Pr Pr\left(a_i,\ b_i,\ \theta_j\right)$$

[34] El modelo de respuesta consiste en modelos logísticos de dos parámetros secuenciales.

Debido a que el modelo de respuesta graduada es básicamente un modelo logístico ordenado, los parámetros de dificultad de cada ítem son naturalmente estimados en un orden creciente. Las dificultades representan un punto en el que una persona con nivel de rasgo latente $\theta_j = b_{ik}$ tiene una probabilidad de 50% de responder la categoría k o una superior. Las estimaciones de los parámetros de dificultad corresponden al punto del continuo del rasgo latente en dónde $\Pr Pr\,(\theta) = 0.5$. Por otro lado, el parámetro de discriminación a_i indica que cuanto mayor es cuan apropiado es para medir el constructor de interés. Valores entre uno y dos indican calidad moderada y valores mayores que dos siguieren alta cualidad.

Usando la forma intercepto-pendiente, la probabilidad de proveer una respuesta k o mejor se parametriza como

$$\Pr Pr\left(a_i,\ b_i,\ \theta_j\right) = \frac{\exp exp\left(\alpha_i\theta_j+\beta_{ik}\right)}{1+\exp exp\left(\alpha_i\theta_j+\beta_{ik}\right)}$$

La transformación de estás dos parametrizaciones es

$$a_i = \alpha_i \qquad b_i = -\frac{\beta_{ik}}{\alpha_i}$$

Sea y_{ij} la respuesta observada para Y_{ij} y $p_{ij} = \Pr Pr\left(\alpha_i,\ \beta_i,\ \theta_j\right)$. Condicional en θ_j, las respuestas a los ítems se asumen independientes de tal manera que la densidad condicional para la persona *j* está dada por

$$f\left(\mathrm{B},\ \theta_j\right) = \prod_{i=1}^{I} p_{ij}$$

Dónde $y_j = \left(y_{1j}, \ldots, y_{Ij}\right)$, $\mathrm{B} = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_I)$, e I se refiere al número de ítems. La verosimilitud para la persona j es calculada al integrar la variable latente de la densidad conjunta

$$L_j(\mathrm{B}) = \int_{-\infty}^{\infty} f\left(\mathrm{B},\ \theta_j\right)\phi(\theta_j)d\theta_j$$

En dónde $\phi(\bullet)$ es la función de densidad para la distribución normal estándar, el logaritmo de la función verosimilitud para la muestra es la suma de los logaritmos de las funciones de verosimilitud de las *N* personas en la muestra.

$$logL(\mathrm{B}) = \sum_{j=1}^{N} logL_j(\mathrm{B})$$

La integral en la fórmula para $L_j(\mathrm{B})$ es generalmente no trazable, por lo que se deben usar métodos numéricos.

El proceso de estimación del modelo de dos parámetros y del modelo de respuesta graduada pasa por dos etapas. En la primera, llamada de calibración, son estimados los parámetros de los ítems y en la segunda, llamada de estimación, son estimadas las medidas de interés. En este caso, para la estimación de los parámetros de ítems se usa el método de máxima verosimilitud marginal. Posteriormente, y en una segunda etapa, para la estimación de las medidas podemos implementar diferentes métodos de estimación de medidas. Entre ellos el más comúnmente usado es el método EAP. Para detalles ver Baker y Kim (2004).

Como indican Bazán, Merino y Mazzon (2011), entre los paquetes comerciales podemos citar IRTPRO, PARSCALE, MULTILOG[35]. También podemos considerar los paquetes ltm y mirt en el programa R[36]. En este estudio es considerado el programa mirt desarrollado por Phil Chalmers (Chalmers, 2012). El programa mirt realiza el proceso en dos las dos etapas ya mencionadas antes, de calibración o de estimación de parámetros de las preguntas y de estimación o específicamente estimación de medidas. El primero se basa en la muestra del estudio de calibración, y usando los valores obtenidos de los parámetros de las preguntas, el proceso de estimación puede ser aplicado a cualquier muestra.

Anexo 4. Calibración de las preguntas

Preguntas cuestionario para primaria

Perseverancia									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p1	3.05	1.01	0.62	0.55	1.849	-2.639	-2.099	-0.984	0.309
P2	2.91	1.06	0.56	0.49	1.242	-3.435	-2.519	-0.973	0.654
p3	3.09	0.95	0.67	0.65	2.598	-2.633	-2.063	-0.997	0.269

Autoeficacia académica									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p4	3.37	0.82	0.71	0.64	2.340	-2.882	-2.471	-1.553	-0.157
p5	3.38	0.78	0.67	0.58	1.756	-3.538	-2.998	-1.795	-0.152
p6	3.38	0.79	0.72	0.67	2.460	-2.941	-2.473	-1.570	-0.171

Intimidad en la amistad									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p7	3.13	1.18	0.75	0.68	2.596	-1.937	-1.507	-1.109	-0.152
p8	2.86	1.32	0.76	0.73	3.277	-1.596	-1.097	-0.739	0.113
p9	2.72	1.35	0.60	0.50	1.444	-2.021	-1.198	-0.623	0.388

Trabajo en equipo

[35] Mayores detalles en http://www.ssicentral.com/irt/.

[36] Mayores detalles en http://cran.r-project.org/web/views/Psychometrics.html

Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p10	2.91	1.08	0.59	0.53	1.470	−2.752	−1.918	−0.812	0.431
p11	3.09	1.01	0.65	0.56	1.629	−2.827	−2.218	−1.235	0.219
p12	2.87	1.17	0.70	0.65	2.047	−2.011	−1.495	−0.670	0.313
p13	2.66	1.19	0.59	0.53	1.504	−2.311	−1.562	−0.389	0.647
p14	2.80	1.16	0.61	0.54	1.582	−2.393	−1.742	−0.675	0.467
p15	3.13	0.96	0.69	0.61	1.975	−2.615	−2.161	−1.091	0.168

Sentido de pertenencia con pares									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p16	2.79	1.22	0.62	0.58	1.488	−2.248	−1.621	−0.675	0.487
p17	2.89	1.13	0.64	0.62	1.814	−2.321	−1.757	−0.707	0.410
p18	3.02	1.12	0.66	0.63	2.282	−2.210	−1.681	−0.874	0.143

Género									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p19	2.81	1.37	0.74	0.65	2.448	−1.612	−1.081	−0.717	0.015
p20	2.46	1.41	0.63	0.56	1.676	−1.679	−0.907	−0.203	0.465
p21	3.19	1.11	0.79	0.69	2.717	−1.928	−1.470	−1.170	−0.183
p22	3.08	1.18	0.73	0.62	2.234	−1.960	−1.529	−1.010	−0.080
p23	2.96	1.24	0.82	0.76	3.078	−1.681	−1.151	−0.713	−0.033
p24	2.82	1.33	0.66	0.56	1.832	−1.836	−1.322	−0.674	0.094
p25	3.35	0.97	0.69	0.53	1.899	−2.627	−2.115	−1.590	−0.355

Regulación emocional							
Pregunta	mean	sd	corr	load	a	b1	b2
p26	1.34	0.76	0.69	0.64	2.019	−1.260	−0.046
p27	1.42	0.73	0.73	0.68	2.231	−1.360	−0.196
p28	1.63	0.67	0.60	0.48	1.342	−2.163	−1.106
p29	1.33	0.79	0.61	0.53	1.493	−1.374	−0.250

Toma de decisiones responsables							
Pregunta	mean	sd	corr	load	a	b1	b2
p30	1.43	0.81	0.58	0.48	1.373	−1.317	−0.603
p31	1.56	0.74	0.75	0.65	2.507	−1.287	−0.720
p32	1.41	0.76	0.68	0.61	1.749	−1.396	−0.309
p33	1.41	0.79	0.68	0.60	1.628	−1.352	−0.409

Empatía							
Pregunta	mean	sd	corr	load	a	b1	b2
p34	1.47	0.69	0.77	0.67	2.504	−1.509	−0.360
p35	1.50	0.68	0.84	0.75	3.249	−1.471	−0.403
p36	1.48	0.69	0.78	0.68	2.407	−1.486	−0.360

p37	1.41	0.70	0.86	0.79	3.520	-1.260	-0.178
p38	1.52	0.67	0.84	0.75	3.152	-1.485	-0.402
p39	1.47	0.68	0.85	0.77	3.446	-1.378	-0.327
p40	1.51	0.67	0.82	0.73	2.855	-1.533	-0.379

Asertividad										
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4	b5
p41	3.47	1.45	0.53	0.61	1.259	-2.302	-1.806	-1.590	-0.991	1.779
p42	3.44	1.59	0.68	0.73	1.967	-1.588	-1.343	-1.160	-0.909	1.136
p43	3.19	1.60	0.64	0.66	1.741	-1.746	-1.365	-1.045	-0.295	1.201
p44	3.43	1.56	0.64	0.70	1.778	-1.763	-1.435	-1.294	-0.652	1.060

Conducta prosocial							
Pregunta	mean	sd	corr	load	a	b1	b2
p46	1.13	0.79	0.57	0.52	1.331	-1.166	0.451
p47	1.38	0.74	0.66	0.63	1.742	-1.510	-0.142
p48	1.23	0.78	0.67	0.66	1.896	-1.113	0.173

Agresión escolar							
Pregunta	mean	sd	corr	load	a	b1	b2
p45	1.15	0.82	0.67	0.60	1.629	-0.858	0.277
p49	1.40	0.77	0.73	0.66	2.198	-1.196	-0.239
p50	1.10	0.86	0.76	0.68	2.096	-0.606	0.253
p51	1.44	0.76	0.67	0.60	1.675	-1.547	-0.407
p52	1.48	0.77	0.73	0.63	2.025	-1.280	-0.535

Acoso escolar						
Pregunta	mean	sd	corr	load	a	b1
p53	0.72	0.45	0.72	0.61	2.208	-0.781
p54	0.64	0.48	0.76	0.63	2.593	-0.440
p55	0.69	0.46	0.65	0.52	1.507	-0.783
p56	0.88	0.33	0.69	0.46	1.734	-1.744
p57	0.90	0.29	0.74	0.47	2.123	-1.804

Diversidad - acciones								
Pregunta	mean	sd	corr	load	a	b1	b2	b3
p58	2.49	0.74	0.76	0.67	2.246	-2.491	-1.771	-0.422
p59	2.43	0.79	0.78	0.70	2.355	-2.346	-1.581	-0.273
p60	2.21	0.93	0.74	0.67	2.170	-1.992	-1.139	0.011
p61	2.43	0.83	0.78	0.70	2.816	-1.967	-1.389	-0.270
p62	2.35	0.85	0.77	0.70	2.632	-1.995	-1.172	-0.105
p63	2.40	0.89	0.75	0.66	2.629	-1.862	-1.249	-0.353

Diversidad - actitudes							
Pregunta	mean	sd	corr	load	a	b1	b2
p64	1.24	0.68	0.71	0.64	1.907	-1.552	0.327
p65	1.18	0.67	0.68	0.60	1.822	-1.428	0.527
p66	1.37	0.69	0.76	0.67	2.628	-1.398	-0.071
p67	1.12	0.72	0.68	0.61	1.869	-1.144	0.539
p68	1.40	0.68	0.79	0.70	2.820	-1.432	-0.109
p69	1.22	0.67	0.71	0.65	2.212	-1.346	0.434
p70	1.18	0.70	0.69	0.62	1.974	-1.275	0.443

Preguntas cuestionario para secundaria

Autopercepción general									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p1	3.16	0.88	0.65	0.62	1.849	-3.219	-2.745	-1.125	0.302
p2	3.02	0.94	0.70	0.70	2.324	-2.655	-2.014	-0.921	0.438
p3	2.86	0.86	0.63	0.61	1.834	-3.077	-2.510	-0.700	0.985

Perseverancia									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p4	2.82	0.93	0.71	0.68	2.016	-2.850	-2.212	-0.552	0.847
p5	2.87	0.86	0.65	0.59	1.709	-3.198	-2.679	-0.706	0.959
p6	2.82	0.90	0.77	0.79	3.675	-2.545	-1.959	-0.429	0.731

Autoeficacia académica									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p7	3.19	0.76	0.71	0.71	2.406	-3.073	-2.462	-1.135	0.386
p8	3.18	0.83	0.61	0.54	1.620	-3.450	-2.693	-1.462	0.399
p9	3.18	0.79	0.70	0.66	2.266	-3.151	-2.600	-1.169	0.391

Intimidad en la amistad									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p10	3.06	1.13	0.83	0.78	4.029	-1.851	-1.313	-0.860	0.080
p11	2.78	1.28	0.85	0.84	4.667	-1.537	-0.986	-0.540	0.272
p12	2.62	1.24	0.70	0.64	1.934	-1.936	-1.108	-0.444	0.632

Trabajo en equipo									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p13	2.89	0.97	0.65	0.60	1.618	-2.792	-2.039	-0.773	0.839
p14	2.88	0.92	0.67	0.63	1.787	-2.986	-2.162	-0.698	0.874
p15	2.69	1.07	0.74	0.70	2.369	-2.065	-1.466	-0.331	0.725

p16	2.60	1.05	0.61	0.56	1.614	-2.620	-1.714	-0.201	1.066
p17	2.62	1.04	0.69	0.65	2.018	-2.305	-1.579	-0.276	0.962
p18	2.94	0.87	0.80	0.76	2.893	-2.542	-1.953	-0.750	0.724

Sentido de pertenencia con pares									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p19	2.69	1.04	0.69	0.67	2.060	-2.389	-1.628	-0.389	0.892
p20	2.84	0.95	0.70	0.68	2.193	-2.638	-1.951	-0.639	0.801
p21	2.87	1.04	0.72	0.70	2.237	-2.392	-1.736	-0.696	0.574

Género									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p22	3.21	1.04	0.74	0.61	2.380	-2.380	-1.753	-1.248	-0.220
p23	3.38	0.88	0.82	0.74	3.212	-2.385	-1.916	-1.243	-0.394
p24	3.50	0.74	0.88	0.80	4.681	-2.349	-1.886	-1.451	-0.453
p25	3.39	0.89	0.81	0.69	3.117	-2.302	-1.813	-1.454	-0.373
p26	3.36	0.91	0.89	0.82	4.534	-2.172	-1.646	-1.147	-0.380
p27	3.24	1.02	0.81	0.72	3.051	-2.191	-1.672	-1.072	-0.304
p28	3.54	0.72	0.79	0.66	2.937	-2.533	-2.232	-1.761	-0.528

Regulación emocional							
Pregunta	mean	sd	corr	load	a	b1	b2
p29	1.13	0.61	0.70	0.63	1.968	-1.546	0.858
p30	1.33	0.73	0.56	0.50	1.226	-1.835	0.036
p31	1.37	0.71	0.73	0.66	2.103	-1.496	-0.041
p32	1.64	0.65	0.59	0.47	1.373	-2.217	-0.966
p33	1.21	0.80	0.52	0.47	1.117	-1.351	0.313
p34	1.12	0.76	0.69	0.62	1.721	-1.037	0.492

Toma de decisiones responsables							
Pregunta	mean	sd	corr	load	a	b1	b2
p35	1.00	0.84	0.76	0.69	2.416	-0.395	0.477
p36	1.06	0.87	0.78	0.70	2.437	-0.423	0.301
p37	0.77	0.77	0.75	0.69	2.324	-0.164	1.022
p38	0.77	0.78	0.73	0.66	2.084	-0.145	1.012

Empatía							
Pregunta	mean	sd	corr	load	a	b1	b2
p39	1.25	0.59	0.76	0.68	2.420	-1.761	0.511
p40	1.27	0.60	0.82	0.74	2.821	-1.609	0.400
p41	1.21	0.66	0.81	0.74	2.588	-1.341	0.395
p42	1.17	0.61	0.85	0.78	3.340	-1.363	0.582
p43	1.33	0.62	0.81	0.74	2.903	-1.631	0.139

p44	1.29	0.62	0.83	0.76	3.190	-1.549	0.263
p45	1.21	0.64	0.79	0.72	2.613	-1.396	0.452

Asertividad										
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4	b5
p46	2.31	1.78	0.65	0.68	1.694	-1.014	-0.309	-0.104	0.507	1.947
p47	2.46	1.86	0.72	0.74	2.037	-0.897	-0.283	-0.169	0.017	1.840
p48	2.29	1.66	0.60	0.61	1.423	-1.152	-0.557	-0.105	0.884	2.353
p49	2.26	1.89	0.74	0.77	2.264	-0.660	-0.132	-0.076	0.391	1.564

Conducta prosocial							
Pregunta	mean	sd	corr	load	a	b1	b2
p51	1.19	0.76	0.64	0.58	1.641	-1.191	0.334
p52	1.47	0.71	0.71	0.65	2.307	-1.455	-0.321
p53	1.26	0.78	0.69	0.65	2.100	-1.090	0.076

Agresión escolar							
Pregunta	mean	sd	corr	load	a	b1	b2
p50	1.13	0.81	0.63	0.56	1.400	-0.949	0.388
p54	1.45	0.76	0.72	0.63	1.912	-1.313	-0.414
p55	1.46	0.76	0.69	0.61	1.812	-1.383	-0.444
p56	1.00	0.85	0.75	0.65	2.046	-0.435	0.522
p57	1.45	0.76	0.70	0.61	1.886	-1.379	-0.435
p58	1.66	0.66	0.72	0.58	2.009	-1.652	-0.963

Acoso escolar						
Pregunta	mean	sd	corr	load	a	b1
p59	0.88	0.32	0.81	0.61	3.014	-1.464
p60	0.81	0.39	0.78	0.58	2.528	-1.134
p61	0.87	0.33	0.77	0.59	2.285	-1.455
p62	0.94	0.24	0.82	0.57	2.997	-1.913
p63	0.94	0.24	0.84	0.61	3.729	-1.809

Diversidad - acciones								
Pregunta	mean	sd	corr	load	a	b1	b2	b3
p64	2.28	0.75	0.70	0.61	1.934	-2.854	-1.483	0.182
p65	2.59	0.66	0.81	0.72	2.722	-2.586	-1.860	-0.601
p66	2.39	0.74	0.74	0.65	1.922	-2.773	-1.679	-0.086
p67	2.64	0.66	0.82	0.72	2.917	-2.586	-1.884	-0.756
p68	2.71	0.61	0.80	0.68	2.618	-2.752	-1.965	-0.918
p69	2.46	0.78	0.75	0.66	2.120	-2.480	-1.679	-0.401
p70	2.68	0.62	0.73	0.62	2.024	-3.071	-2.234	-1.003
p71	2.69	0.62	0.83	0.72	2.917	-2.554	-1.900	-0.871

Diversidad - actitudes							
Pregunta	mean	sd	corr	load	a	b1	b2
p72	1.30	0.56	0.72	0.64	1.984	-2.228	0.489
p73	1.22	0.55	0.72	0.65	2.054	-1.959	0.697
p74	1.33	0.58	0.81	0.74	2.869	-1.833	0.279
p75	1.16	0.59	0.71	0.64	1.928	-1.682	0.879
p76	1.43	0.60	0.81	0.72	3.046	-1.818	-0.026
p77	1.23	0.57	0.75	0.69	2.281	-1.874	0.614
p78	1.22	0.56	0.79	0.74	2.658	-1.742	0.620

Preguntas cuestionario para media

Autopercepción general									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p1	3.09	0.89	0.68	0.65	1.861	-3.341	-2.689	-0.872	0.503
p2	3.03	0.92	0.74	0.74	2.964	-2.871	-1.890	-0.799	0.523
p3	2.92	0.83	0.68	0.66	1.848	-3.637	-2.665	-0.771	0.945

Proyecto de vida									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p4	3.04	0.96	0.74	0.69	2.854	-2.579	-1.794	-0.697	0.383
p5	3.05	0.89	0.77	0.72	3.341	-2.634	-1.849	-0.774	0.450
p6	3.09	0.81	0.77	0.72	2.016	-3.396	-2.462	-1.034	0.618
p7	2.99	0.84	0.60	0.51	0.969	-6.131	-4.055	-0.980	1.315
p8	3.28	0.76	0.66	0.58	1.453	-4.384	-3.077	-1.232	0.211

Perseverancia									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p9	2.78	0.87	0.75	0.72	2.460	-3.082	-2.226	-0.384	1.060
p10	2.92	0.79	0.68	0.63	1.796	-3.738	-2.691	-0.552	0.950
p11	2.82	0.82	0.82	0.84	6.062	-2.719	-1.938	-0.317	0.861

Autoeficacia académica									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p12	3.31	0.77	0.68	0.59	1.951	-3.645	-2.764	-1.610	0.067
p13	3.15	0.76	0.74	0.70	2.365	-3.090	-2.529	-1.061	0.485
p14	3.12	0.79	0.62	0.54	1.333	-4.533	-3.101	-1.519	0.649
p15	3.14	0.75	0.73	0.68	2.033	-3.679	-2.370	-0.664	0.536
p16	3.64	0.59	0.80	0.67	3.357	-2.935	-2.086	-1.273	-0.591

Intimidad en la amistad									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4

p17	2.94	1.16	0.88	0.84	5.216	-1.702	-1.272	-0.732	0.164
p18	2.75	1.25	0.90	0.91	7.725	-1.547	-1.029	-0.510	0.253
p19	2.57	1.21	0.70	0.62	2.047	-2.035	-1.218	-0.365	0.676

Trabajo en equipo									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p20	2.89	0.93	0.62	0.58	1.667	-3.157	-2.247	-0.867	0.836
p21	2.88	0.87	0.72	0.68	1.942	-3.050	-2.317	-0.779	0.934
p22	2.59	1.02	0.77	0.73	2.703	-2.024	-1.357	-0.122	0.973
p23	2.38	1.06	0.55	0.50	1.552	-2.487	-1.416	0.158	1.503
p24	2.59	0.91	0.73	0.70	2.179	-2.695	-1.720	-0.095	1.253
p25	2.91	0.82	0.79	0.75	2.711	-2.818	-2.272	-0.791	0.856

Sentido de pertenencia con pares									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p26	2.60	0.95	0.71	0.69	2.129	-2.634	-1.810	-0.189	1.288
p27	2.83	0.91	0.76	0.76	3.024	-2.476	-1.766	-0.584	0.809
p28	2.81	0.98	0.71	0.69	2.211	-2.517	-1.870	-0.656	0.803

Género									
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4
p29	3.41	0.88	0.77	0.64	2.531	-2.733	-2.103	-1.600	-0.453
p30	3.42	0.86	0.79	0.68	2.856	-2.565	-2.175	-1.438	-0.531
p31	3.60	0.65	0.89	0.80	5.069	-2.361	-2.076	-1.563	-0.659
p32	3.56	0.70	0.83	0.70	3.583	-2.600	-2.129	-1.728	-0.631
p33	3.52	0.75	0.88	0.80	4.446	-2.510	-1.998	-1.423	-0.628
p34	3.40	0.92	0.83	0.72	3.353	-2.250	-1.823	-1.401	-0.561
p35	3.62	0.62	0.82	0.68	3.238	-2.813	-2.271	-1.514	-0.677

Regulación emocional							
Pregunta	mean	sd	corr	load	a	b1	b2
p36	1.17	0.60	0.71	0.63	2.046	-1.665	0.881
p37	1.30	0.69	-0.47	0.40	0.930	-2.406	0.468
p38	1.53	0.66	0.52	0.45	1.088	-2.796	-0.705
p39	1.39	0.68	0.75	0.67	2.112	-1.688	0.069
p40	1.69	0.60	0.53	0.39	0.982	-3.182	-1.405
p41	1.13	0.80	0.53	0.47	0.927	-1.329	0.606
p42	1.15	0.73	0.67	0.62	1.498	-1.336	0.611

Toma de decisiones responsables							
Pregunta	mean	sd	corr	load	a	b1	b2
p43	0.81	0.79	0.79	0.72	2.849	-0.108	0.969
p44	0.80	0.81	0.79	0.72	2.755	-0.039	0.918

p45	0.64	0.70	0.81	0.74	2.504	-0.001	1.372
p46	0.54	0.67	0.76	0.66	2.010	0.232	1.791

Seguimiento de la ley							
Pregunta	mean	sd	corr	load	a	b1	b2
p47	1.67	0.62	0.73	0.60	2.080	-2.123	-1.006
p48	1.57	0.66	0.69	0.57	1.720	-2.137	-0.723
p49	1.68	0.64	0.76	0.62	2.140	-1.859	-1.121
p50	1.67	0.63	0.82	0.72	3.289	-1.730	-0.919

Empatía							
Pregunta	mean	sd	corr	load	a	b1	b2
p51	1.16	0.46	0.77	0.64	2.390	-2.214	0.883
p52	1.21	0.49	0.82	0.72	2.742	-2.135	0.635
p53	1.16	0.56	0.80	0.73	2.897	-1.539	0.600
p54	1.13	0.50	0.87	0.79	4.230	-1.554	0.731
p55	1.27	0.55	0.85	0.77	3.773	-1.805	0.310
p56	1.25	0.56	0.90	0.83	6.908	-1.511	0.297
p57	1.17	0.57	0.78	0.71	2.279	-1.625	0.650

Asertividad										
Pregunta	mean	sd	corr	load	a	b1	b2	b3	b4	b5
p58	2.26	1.84	0.68	0.69	1.790	-0.923	-0.252	-0.106	0.636	1.314
p59	2.32	1.94	0.71	0.71	1.942	-0.773	-0.101	-0.043	0.127	1.372
p60	2.27	1.60	0.59	0.58	1.478	-1.244	-0.669	-0.118	1.202	1.747
p61	2.10	2.00	0.73	0.73	2.242	-0.463	0.045	0.095	0.517	1.089

Agresión escolar							
Pregunta	mean	sd	corr	load	a	b1	b2
p63	1.20	0.76	0.60	0.53	1.291	-1.458	0.268
p64	1.61	0.65	0.76	0.70	3.936	-1.686	-0.753
p65	1.30	0.79	0.71	0.65	1.597	-1.290	-0.049

Conducta prosocial							
Pregunta	mean	sd	corr	load	a	b1	b2
p62	1.23	0.78	0.67	0.59	1.551	-1.154	0.206
p66	1.33	0.81	0.63	0.55	1.406	-1.224	-0.183
p67	1.54	0.71	0.84	0.74	2.865	-1.456	-0.591
p68	0.95	0.84	0.72	0.61	1.868	-0.279	0.666
p69	1.53	0.73	0.73	0.63	1.780	-1.526	-0.644
p70	1.76	0.55	0.71	0.54	1.863	-2.170	-1.389

Acoso escolar

Pregunta	mean	sd	corr	load	a	b1
p71	0.95	0.23	0.86	0.64	3.564	-1.869
p72	0.88	0.33	0.82	0.57	4.254	-1.317
p73	0.93	0.25	0.81	0.59	3.629	-1.728
p74	0.97	0.17	0.84	0.56	3.961	-2.269
p75	0.97	0.17	0.87	0.60	5.302	-2.320

Diversidad - acciones								
Pregunta	mean	sd	corr	load	a	b1	b2	b3
p76	2.47	0.65	0.70	0.57	2.020	-3.219	-1.651	-0.154
p77	2.75	0.52	0.83	0.70	2.991	-3.075	-2.079	-0.936
p78	2.39	0.70	0.66	0.54	1.604	-3.433	-1.877	-0.006
p79	2.74	0.56	0.82	0.70	2.537	-3.157	-1.955	-0.939
p80	2.79	0.49	0.75	0.59	1.892	-3.717	-2.345	-1.213
p81	2.62	0.65	0.74	0.63	1.796	-3.411	-1.870	-0.625
p82	2.83	0.47	0.80	0.65	2.521	-3.164	-2.444	-1.492
p83	2.84	0.43	0.86	0.70	3.452	-2.973	-2.254	-1.245

Diversidad - actitudes							
Pregunta	mean	sd	corr	load	a	b1	b2
p84	1.36	0.52	0.80	0.71	2.664	-2.346	0.290
p85	1.24	0.50	0.71	0.64	1.931	-2.519	0.748
p86	1.39	0.54	0.84	0.77	3.727	-2.018	0.106
p87	1.21	0.55	0.72	0.64	1.968	-1.878	0.758
p88	1.46	0.55	0.82	0.72	3.265	-2.056	-0.120
p89	1.28	0.53	0.78	0.71	2.592	-2.083	0.480
p90	1.30	0.52	0.88	0.82	3.571	-1.894	0.442