DESARROLLO DE UNA EVALUACIÓN DE DIAGNÓSTICO COGNITIVO DE PENSAMIENTO ESTADÍSTICO CON EL MÉTODO RULE SPACE

(Tesis de Maestría en Psicología)

Victor Alexander Rivera Mancilla

Directora: Aura Nidia Herrera Rojas

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

Línea de investigación: Psicología Básica y Experimental

Sublínea: Métodos e Instrumentos de Investigación en Ciencias del Comportamiento

Y entonces, ¿Qué es el hombre, por sí mismo, sino un insecto fútil que zumba mientras se estrella contra el cristal de una ventana? Y es que está ciego, no puede ver, ni puede darse cuenta de que hay algo entre él y la luz. Por eso se esfuerza, trabajosamente, en acercarse. Puede apartarse de la luz, pero no es capaz de llegar a estar más cerca. ¿Cómo le ayudará la ciencia?

Fernando Pessoa (Diarios)

Para mi familia: a mis padres, Victor Manuel y Luz Mélida, que me cuidaron, me dieron amor, me enseñaron a tener paciencia y a hacer las cosas bien. Para mis hermanas, Cristina y Diana, y mis sobrinos Jerónimo, Martín y Simone, que me recuerdan el valor de la incondicionalidad y siempre me animan a dar pasos más largos. Para Aliz, con quien tengo la dicha de la amistad sincera y el amor vivo, y a toda su familia que es la mía también.

Para mis amigos: especialmente aquellos de toda la vida Leonardo y Hollman con quienes comparto mis alegrías y no tantas penas, y a sus familias. A mis amigos y compañeros de la Universidad Nacional de Colombia, que me enseñaron a buscar la excelencia académica, a dudar, debatir y sobresalir.

AGRADECIMIENTOS

"*Una mano más una mano no son dos manos; Son manos unidas…*"

Gonzalo Arango

Elaborar este trabajo no hubiera sido posible sin la ayuda de otras manos, varias manos. Por eso quiero dejar plasmado mi agradecimiento:

A la profesora Aura Nidia Herrera Rojas, quien me ha enseñado y guiado en la elaboración de este trabajo. Es ejemplo de rigurosidad y calidad académica, y también de compromiso social.

A mis compañeros del grupo de investigación en Métodos e Instrumentos de Investigación en Ciencias del Comportamiento, que con sus comentarios y sugerencias ayudaron a mejorar el trabajo. A mis compañeros de cohorte Ana Cristina Santana, Jazmine Escobar y especialmente a Ricardo Narváez, que contribuyeron en la formulación del Proyecto de Tesis.

A los estudiantes, docentes y directivos docentes de Valledupar y La Paz (Cesar), Tumaco (Nariño), San José del Guaviare y Bogotá. Los que hicieron posible este estudio, posibilitando la recolección de datos y también realizando observaciones importantes.

A la Universidad de Costa Rica y los docentes que me acogieron en la pasantía que realicé allí y que sirvió para intercambiar conocimientos y experiencias valiosas. A Costa Rica, hermoso lugar que me inspiró para continuar mi trabajo. A todas las personas que conocí allí, especialmente a la señora Argentina Medrano que me entregó una bonita amistad.

A los jurados de este trabajo. El profesor Mario Córdoba que desde su especialidad me recomendó mejorar aspectos importantes del trabajo. Al profesor Álvaro Artavia cuyo aporte a este trabajo es invaluable gracias a sus recomendaciones y a las herramientas que me facilitó para el uso del *Rule Space.*

A mis compañeros del Laboratorio de Psicometría de la Universidad Nacional de Colombia, que son una familia y han hecho que el Laboratorio sea un referente a nivel local, pro también nacional.

Al Departamento de Psicología de la Universidad Nacional de Colombia, a su cuerpo docente, uno de los mejores del país, donde se aprende psicología integralmente, pero también valores.

A mi querida Universidad Nacional de Colombia. Donde tuve la fortuna de formarme para servirle a los míos y al país.

CONTENIDO

Resumen

Las dinámicas educativas actuales demandan que la evaluación ofrezca información útil para mejorar los procesos de enseñanza y aprendizaje en tiempo real. Las Evaluaciones de Diagnóstico Cognitivo entregan información detallada sobre el desempeño de los estudiantes, aprovechando las bondades de la unión entre la psicometría y la psicología cognitiva. Este tipo de evaluaciones pueden ser especialmente útiles en las áreas STEM, relevantes para el desarrollo de los países. Este trabajo tuvo como objetivo principal implementar una Evaluación de Diagnóstico Cognitivo del Pensamiento Estadístico en estudiantes de primer ciclo de básica primaria utilizando el método *Rule Space*. Para ello fue necesario desarrollar una prueba que fue aplicada a 1580 estudiantes de grado segundo a quinto de primaria de cinco diferentes ciudades de Colombia. Los resultados muestran que la prueba es confiable y presenta algunas evidencias de validez adecuadas. Además, los resultados de la aplicación del método *Rule Space* son satisfactorios debido a que logra clasificar a más del 90% de los estudiantes en estados de conocimiento con significado educativo.

Palabras clave: Método Rule Space, Evaluaciones de Diagnóstico Cognitivo, Modelos de Diagnóstico Cognitivo, Evaluación del Pensamiento Estadístico, Alfabetismo Estadístico.

Abstract

Current educational trends demand useful information from educational assessment in order to improve the teaching and learning processes. Cognitively Diagnostic Assessments provide detailed information about examinee´s performance, linking psychometrics and cognitive psychology. These assessments can be very useful in key areas such STEM, that are relevant to the development of nations. The aim of this study was to use a Cognitively Diagnostic Assessment of Statistical Thinking in elementary students using the *Rule Space* method. Was necessary to develop a test that was presented to 1580 students from five Colombian cities. The results showed test reliability and some validity evidences was reached. In the other hand, the *Rule Space* showed a classification rate above 90%. This result implies that the proposed cognitive model about statistical thinking in this sample was acceptable.

Keywords: Rule Space Method, Cognitively Diagnostic Assessment, Cognitively Diagnosis Models, Statistical Thinking Assessment, Statistical Literacy.

Introducción

A lo largo de la historia, la ciencia y la tecnología han propiciado el mejoramiento de la calidad de vida de la gente, especialmente en los países donde se han presentado los mayores desarrollos en estas áreas. Es por eso que en la actualidad existe una exigencia sobre la necesidad de fortalecer las materias que se han agrupado en el término STEM (en inglés, por las iniciales de ciencia, tecnología, ingeniería y matemáticas), especialmente en el plano de la educación. Esto se advierte al conocerse los esfuerzos que diversos gobiernos del mundo han realizado para fortalecer la enseñanza y el aprendizaje de estas áreas consideradas clave para el desarrollo social y económico a mediano y largo plazo. Por ejemplo, desde el gobierno de los Estados Unidos se han realizado esfuerzos importantes a nivel de política pública para aumentar la competencia de los estudiantes en estas áreas (Gonzalez & Kuenzi, 2012; White House Office of Science and Technology Policy, 2014). Incluso, a nivel laboral las dinámicas actuales están demandando profesionales y técnicos altamente capacitados en áreas STEM, por lo que la necesidad de fortalecer la instrucción en estos campos de conocimiento se ha convertido en una prioridad para los países que se quieren mantener a la vanguardia. Sin embargo, es importante tener en cuenta que, aunque como sigla STEM representa cuatro áreas específicas (Ciencia, Tecnología, Ingeniería y Matemáticas), existen definiciones que incluyen otras como la psicología, la economía, la estadística, etc. En trabajos como el de Gonzalez y Kuenzi (2012) se aclara que en la actualidad existen límites difusos entre las áreas de conocimiento en el quehacer académico actual y que, como tal, otras áreas pueden hacer parte de STEM.

Evaluación de Pensamiento Estadístico con *Rule Space*

El Programa para la Evaluación Internacional de Alumnos (PISA, por sus siglas en inglés) de la Organización para la Cooperación y el Desarrollo Económico (OCDE) ha servido para evaluar la calidad de la educación de los países en áreas STEM como las matemáticas y otras como la competencia lectora y el conocimiento en ciencias naturales. De hecho, el enfoque de la prueba ha propiciado la formulación o modificación de políticas públicas a nivel educativo en muchos países al descubrir que existe una relación directa entre los puntajes agregados PISA de los países y su ingreso per cápita (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura - UNESCO, 2015). Es por eso que el desempeño de los estudiantes en estas áreas clave puede ser tomado como un indicador del estado actual y de la proyección de desarrollo económico.

Colombia ha participado en esta evaluación; en la aplicación del año 2012, que tuvo como énfasis las matemáticas, los resultados analizados y reportados por el Instituto Colombiano para la Evaluación de la Educación - ICFES (2013) muestran que el país se ubicó en el puesto 62 entre 65 naciones en el área mencionada. En la aplicación del 2015 ocupó el mismo puesto, pero entre 72 naciones evaluadas (ICFES, 2016). Según estos informes los estudiantes colombianos de 15 años se encuentran, en su mayoría, en el nivel más bajo en cuanto a la habilidad matemática (74% en el 2012 y 66% en el 2015), siendo ésta el área de peor desempeño entre las tres evaluadas. Los resultados de Colombia siguen estando muy lejos del promedio de los países de la OCDE y están entre los más bajos de los países latinoamericanos. Entonces se evidencian falencias que, en el caso específico de las matemáticas, plantean desafíos para todo el sistema educativo colombiano.

Estos desafíos deben ser asumidos por toda la comunidad educativa, donde aquellos que trabajan en el campo de la evaluación tienen mucho que aportar. Asumiendo que la educación debe asegurar un nivel aceptable de comprensión por parte de los estudiantes, se espera que la evaluación sea capaz de dar cuenta de los procesos y habilidades que utilizan los estudiantes para poder decir que entienden o dominan un contenido específico, pero también que pueda brindar información sobre las dificultades que tienen aquellos que aún no lo han logrado. Ejecutar procesos evaluativos que estén relacionados directamente con lo que se realiza en el aula y que sirvan como herramientas para el trabajo de los docentes es un imperativo profesional para los expertos de la evaluación educativa, pero también es un mandato ético (Norris, Macnab & Phillips, 2007).

De hecho, la necesidad de diseñar y desarrollar evaluaciones educativas informativas es una demanda de la educación del siglo XXI (Pellegrino, 2014). Según Huff y Goodman (2007), los docentes esperan que las evaluaciones no solo entreguen información general sobre el desempeño de los estudiantes, sino que también se pueda retroalimentar el proceso de enseñanza y así potenciar el aprendizaje.

Pellegrino, Chudowsky y Glaser (2001), indican que el diseño y el desarrollo de evaluaciones educativas debe ser un proceso de razonamiento a partir de la evidencia y proponen el *triángulo de la evaluación* como marco conceptual para cumplir con esta tarea sin importar el tipo de evaluación que se pretenda llevar a cabo. Los tres componentes del triángulo representan elementos fundamentales que se debe tener en cuenta al momento de diseñar y ejecutar cualquier evaluación educativa: un modelo teórico (preferiblemente

sustentado empíricamente) sobre la *cognición* del estudiante, una serie de consideraciones sobre las *observaciones* (o tareas) que entregarán información sobre el desempeño de los estudiantes y las *interpretaciones* realizadas al final del proceso (teniendo en cuenta la evidencia).

Los elementos del triángulo no siempre son explícitos, pero deben considerarse. También, el triángulo implica que los tres elementos están conectados y son interdependientes. El diseño y desarrollo de evaluaciones se convierte en un proceso iterativo en donde cada uno de los tres vértices retroalimenta a los demás. Sin embargo, el punto de partida se encuentra en la esquina de la *cognición*: en la medida en que se plantee con claridad el modelo de aprendizaje que va a sustentar el resto del diseño de la evaluación, esta podrá ser más efectiva en términos educativos.

Las evaluaciones que parten desde modelos cognitivos robustos y que al final entregan información diagnóstica útil son llamadas Evaluaciones de Diagnóstico Cognitivo (en adelante también EDC). Huff y Goodman (2007) realizan una diferenciación entre las evaluaciones psicométricas y las EDC, señalando que las primeras tienen un objetivo principalmente descriptivo, mientras que las otras buscan establecer un perfil detallado de los estados de aprendizaje de los estudiantes y ofrecer información para corregir o mejorar. La implementación exitosa de las EDC requiere la comunión entre la psicología cognitiva y la psicometría.

No obstante, el vínculo de las pruebas con el trabajo que se realiza en el aula e incluso con los procesos de aprendizaje de los estudiantes no ha sido claro. Por ejemplo, Borsboom (2006)

reconoce que ha existido un distanciamiento entre la psicometría y la psicología, reflejado entre otras cosas, en el hecho de que la teoría psicológica no ha motivado modelos psicométricos específicos. Por otra parte, Leighton y Gierl (2007) mencionan que la psicología cognitiva puede aportar elementos importantes para diseñar pruebas en el campo educativo que cumplan una función sumativa y a la vez formativa, basadas en teorías con fundamento empírico.

Referentes Conceptuales

Pensamiento Estadístico

La estadística se ha ganado un lugar importante en el campo de las ciencias en general, pero también es importante en la vida cotidiana de las personas. En el caso de la ciencia, la estadística aporta métodos y procedimientos importantes para el desarrollo de las investigaciones en diversas áreas de conocimiento, a tal punto que la producción científica sobre estadística puede encontrarse en revistas o libros de disciplinas como la psicología, la economía, las matemáticas, etc. De hecho, existen autores que afirman que las habilidades en estadística son casi imprescindibles para los científicos actuales (Cox & Efron, 2017); otros mencionan que el "Pensamiento Estadístico" ayuda al éxito y a la calidad de las conclusiones de cualquier estudio (Tong, 2019). En la vida de las personas, la utilidad de la estadística es menos evidente, pero existen investigaciones que muestran la relación del dominio de conceptos estadísticos básicos con la capacidad para tomar decisiones en contextos políticos (e.g., Arnold, 2017), con la evaluación de afirmaciones sobre información factual (e.g., Engel, 2017), con la valoración de evidencia médica sobre riesgos en salud (e.g., Wegwarth & Gigerenzer, 2018), con el desarrollo del pensamiento crítico (e.g., Aizikovitsh-Udi , Kuntze & Clarke, 2016; Ben-Zvi & Makar, 2016; Kuntze, Aizikovitsh-Udi & Clarke, 2017), etc. Esto posiciona a la educación estadística como una de las áreas con mayor atención en la escuela actualmente.

Lo anterior haría suponer que todos los estudiantes deberían ser competentes en el área de la estadística, incluso a un nivel informal (Gal, 2002). Sin embargo, esto no es así: muchos estudiantes muestran una ansiedad hacia la estadística y los docentes deben realizar grandes esfuerzos para que los estudiantes logren un desempeño adecuado en esta área (Ben-Zvi & Makar, 2016). A pesar de la gran atención que ha recibido la educación estadística en muchos países, los problemas respecto al razonamiento inapropiado en esta área se mantienen (Garfield & Ben-Zvi, 2007).

Según Ben-Zvi & Garfield (2004), el área de la investigación en la educación estadística históricamente ha tenido el problema de que las definiciones no han sido claras ni unificadas. Durante el siglo XX muchos autores hablaron de conceptos como *alfabetismo estadístico, pensamiento estadístico, conocimiento estadístico, razonamiento estadístico* indistintamente y sin precisar a qué se referían. Recientemente, se ha tratado de corregir el hecho de que incluso en eventos especializados sobre educación estadística no había consenso sobre términos básicos entre los investigadores. Solo hasta el final del Siglo XX e inicio del XXI fue cuando se trató de establecer definiciones claras al respecto; por ejemplo, Moore (1997) indicó que el *pensamiento estadístico* está compuesto por: la necesidad de datos, la importancia de la producción de datos, la omnipresencia de la variabilidad, el concepto de medida y la modelación de la variabilidad.

Después, Wild y Pfannkuch (1999) plantearon un marco más amplio para entender el *pensamiento estadístico* en el proceso de investigación científica, compuesto por cinco dimensiones: a) El ciclo investigativo (compuesto a su vez por cinco elementos: Problema,

Plan, Datos, Análisis y Conclusiones), b) Tipos de pensamiento (Tipos generales: Estratégico, búsqueda de explicaciones modelación y aplicación de técnicas, y los Tipos fundamentales del pensamiento estadístico: Reconocimiento de la necesidad de datos, transnumeración, consideración de la variación, razonamiento mediante modelos estadísticos e integración de lo estadístico y lo contextual), c) El ciclo interrogativo (Generar, Indagar, Interpretar, Criticar y Juzgar), y d) Las disposiciones (Escepticismo, Imaginación, Curiosidad y conciencia, Apertura, Disposición para indagar de manera profunda, Ser lógico, Compromiso y Perseverancia). Estos autores, en su propuesta también enfatizan sobre la importancia del contexto en cualquier investigación que se apoye en el *pensamiento estadístico*. Esta propuesta fue importante debido a que abrió el camino para pensar en una definición del pensamiento estadístico, sobre su utilidad en el contexto investigativo, pero también en la cotidianidad:

> *We recognise that much statistical thinking can beneficially take place in day-to-day activities, particularly in the interpretation of information in media and other reports. In interpreting reports, we recognise the applicability of parts of our statistical knowledge about the production, behaviour and analysis of data to the type of information we are receiving and are thus able to critically appraise aspects of that information. p 224.*

Pese a que a Wild y Pfannkuch (1999) se les critica por proveer un marco de referencia del pensamiento estadístico bastante amplio (que podría ser equiparable al concepto de *pensamiento científico*), se reconoce su intención de responder a la pregunta ¿qué es el pensamiento estadístico?, lo cual fue el antecedente para que se sintetizaran las discusiones y los consensos al respecto planteados en diferentes encuentros académicos en el campo de

la educación estadística. El modelo más aceptado y difundido en la actualidad sobre la organización del *pensamiento estadístico* es el propuesto por Ben-Zvi y Garfield (2004) quienes brindan claridad a la definición de este concepto y presentan otras relacionadas para diferenciar entre alfabetismo, razonamiento y pensamiento estadísticos:

- Alfabetismo estadístico (*Statistical literacy*): Aquí se incluyen las habilidades básicas que sirven para entender información estadística. Implica habilidades para comprender conceptos fundamentales de la estadística, vocabulario y símbolos. Se incluye también la capacidad para entender lo que significa el concepto de probabilidad y la noción de incertidumbre y variabilidad.
- Razonamiento estadístico (*Statistical reasoning*): Este es un paso más allá respecto al alfabetismo estadístico y está definido como la forma en que la persona le da sentido a la información estadística, también tiene que ver con la manera en que las personas procesan la información proveniente de tablas, gráficos y porcentajes. Implica entender y poder explicar los procedimientos estadísticos para ser capaz de interpretar completamente información estadística.

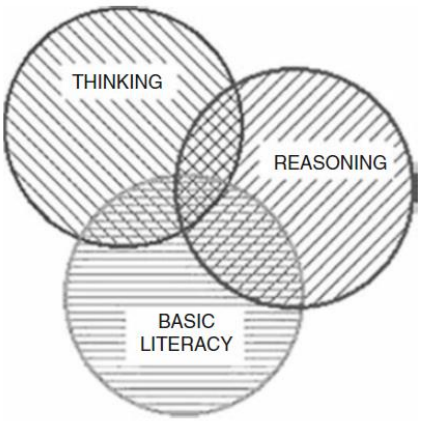

Figura 1. Representación gráfica de la jerarquía entre alfabetismo, razonamiento y pensamiento estadístico. Tomado de Garfield (2011).

- Pensamiento estadístico (*Statistical thinking*): Tiene que ver con las grandes ideas y el entendimiento completo (a nivel profesional, pero no exclusivo de los profesionales en estadística) de cómo las investigaciones estadísticas son conducidas. Implica el entendimiento adecuado de conceptos como la variación, la probabilidad, las medidas de tendencia central, el conocimiento pleno de cómo realizar una investigación y los límites de la inferencia a partir de datos. También, las personas que dominan pensamiento estadístico son capaces de realizar evaluaciones y críticas sobre la forma en que los demás dan tratamiento a la información.

Garfield (2011) indica que cada uno de estos tres conceptos involucran resultados de aprendizaje independientes, con una estructura jerárquica pero que se solapan entre sí, como se muestra en la Figura 1. El alfabetismo estadístico es sobre el que se fundamentan los otros dos conceptos y el pensamiento estadístico es de orden superior respecto a los demás.

Evaluación de Pensamiento Estadístico con *Rule Space*

Garfield (2011) también señala que una forma de distinguir entre los tres conceptos expuestos arriba es observando el tipo de palabras que se usan para denotar la forma en que se evalúa cada uno de ellos. Estas palabras se relacionan con los procesos propuestos en la taxonomía de Bloom (1956; citado por Garfield, 2011) y se muestran en la Tabla 1. De acuerdo con ello se reafirma la idea de una jerarquía, teniendo en cuenta entre otras cosas, que las palabras utilizadas para evaluar el pensamiento estadístico son aquellas que aparecen en lo más alto de la pirámide de los objetivos educativos establecidos por Bloom.

La jerarquía propuesta por Ben-Zvi y Garfield (2004) describe de manera amplia diferentes niveles de manejo de información estadística, pero que no todas las personas llegan a dominar. Por ejemplo, Tarr y Lannin (2005) mencionan que los estudiantes de educación media presentan bastantes dificultades respecto al tema de la estadística y que persisten incluso en el nivel universitario; entre ellas están problemas en las nociones de condicionalidad y causación, la falacia del eje del tiempo, dificultad en el manejo del concepto de probabilidad de eventos independientes, etc. También se ha encontrado que la mayoría de los estudiantes universitarios muestran bastantes vacíos a la hora de comprender conceptos estadísticos, incluso después de haber tomado y aprobado un curso de estadística (Zieffler, Garfield, Alt, Dupuis, Holleque & Chang, 2008). Es claro entonces, que lograr que los estudiantes tengan un buen dominio de temas estadísticos es un reto, por lo que los esfuerzos deberían ser grandes para lograr un cambio al respecto, desde el inicio mismo de la formación y comenzando con la parte baja de la jerarquía: el *alfabetismo estadístico*.

Tabla 1

Verbos Utilizados Para Evaluar Alfabetismo, Razonamiento y Pensamiento Estadísticos.

Alfabetismo	Razonamiento	Pensamiento
Identificar, Describir, Traducir, Interpretar, Leer, Operar.	Explicar por qué, Explicar cómo.	Aplicar, Criticar, Evaluar, Generalizar

Nota. Tomado de Garfield (2011).

Watson y Callingham (2003, 2005) propusieron una jerarquía para el alfabetismo estadístico en estudiantes de colegio. La jerarquía propuesta, que se muestra en la Tabla 2, consta de 6 niveles que van desde un nivel de entendimiento *idiosincrático* asociado a capacidades básicas de manejo de información como conteo e identificación local en tablas, hasta el nivel *Crítico matemático* en donde precisamente se utiliza un razonamiento crítico, se usa también un razonamiento basado en proporciones en tareas que involucran el azar, se reconoce el concepto de la incertidumbre y se es capaz de interpretar diferencias sutiles en el lenguaje en las tareas que se les presentan. Este modelo se caracteriza por asumir una sola dimensión del alfabetismo estadístico, con seis niveles y ha servido para desarrollar evaluaciones de este dominio. Este modelo fue construido empíricamente, utilizando las respuestas de miles de estudiantes de Australia a diferentes ítems diseñados a partir de lineamientos curriculares referentes al manejo de datos y de azar o probabilidad y clasificando los ítems en grupos de acuerdo con su dificultad y a las exigencias de cada uno de ellos.

Tabla 2

Jerarquía propuesta por Callingham y Watson (2017) para el alfabetismo estadístico.

Nivel	Nombre	Características del nivel
1	Idiosincrático	Relación idiosincrática con el contexto, uso tautológico de términos y habilidades matemáticas básicas asociadas al conteo y a la identificación de información en celdas en tablas.
2	Informal	Relación coloquial o informal con el contexto indicando creencias no-estadísticas, consideración de elementos aislados dentro de ambientes complejos y capacidad para hacer cálculos de un solo paso en tablas, gráficos y situaciones de azar.
3	Inconsistente	Relación selectiva con el contexto, reconocimiento de conclusiones adecuadas sobre los datos, pero sin ofrecer justificación y privilegio del uso cualitativo de las ideas estadísticas.
4	Consistente No Crítico	Relación apropiada con el contexto, pero no crítico, uso de terminología adecuado, apreciación de la variabilidad solo en contextos de azar y habilidades estadísticas asociadas a la media, probabilidad simple y características de gráficas.
5	Crítico	Crítico, uso apropiado de la terminología estadística, interpretación cualitativa del azar y apreciación de la variabilidad.
6	Crítico matemático	Crítico, relacionado con el contexto, uso de razonamiento sobre proporcionalidad en tareas con azar, muestra de la necesidad de la incertidumbre cuando se hacen predicciones e interpretación apropiada de sutilezas en el lenguaje.

Nota. Tomado de Callingham y Watson (2017). Traducción propia.

A su vez, Callingham y Watson (2017) evalúan su propuesta después de más de una década, reafirmando la jerarquía sugerida en los estudios anteriores, manteniendo su definición del alfabetismo estadístico como constructo unitario, pero sin especificar su relación con propuestas como la de Ben-Zvi y Garfield (2004).

Es claro que las habilidades mencionadas como integrantes del pensamiento, razonamiento y alfabetismo estadístico son muy importantes para el desempeño de los estudiantes desde los

primeros grados hasta la universidad, pero también en lo cotidiano. Es por eso que muchos sistemas educativos han integrado la estadística y la probabilidad dentro de los currículos en todos los niveles y los que ya lo habían hecho, están enfatizando en ello. Calllingham y Watson (2017) mencionan dos ejemplos: el currículo en Australia cambió la parte llamada "Azar y datos" por "Estadística y probabilidad", mostrando la necesidad de refinar los contenidos incluidos; y el caso de Nueva Zelanda en donde el alfabetismo estadístico fue incluido como una subsección específica en el currículo de matemáticas y estadística.

Por último, Gafield & Ben-Zbi (2008) mencionan que muchos investigadores en esta área, más allá de enfocarse en las definiciones de conceptos como pensamiento estadístico, prefieren concentrarse en las "grandes ideas" de la estadística y estudiar cómo lograr que los estudiantes las dominen. Según estos autores estas ideas son presentadas a los estudiantes dentro de todo el proceso educativo, pero pocos logran dominarlas, por lo que los educadores deben esforzarse para que los estudiantes logren aprehenderlas en lugar de solamente dominar habilidades asiladas (e.g., saber cómo hallar la mediana). Estas grandes ideas son: datos, distribución, tendencia, variabilidad, modelo, asociación, muestra y muestreo e inferencia.

En Colombia, los Estándares Básicos de Competencias (En adelante "EBC" o simplemente "Estándares") del Ministerio de Educación Nacional (2006) para el área de Matemáticas establecen los conocimientos y competencias fundamentales en términos de "habilidades" o destrezas que todos los estudiantes debe dominar. Allí se propone el concepto de pensamiento matemático dividido en cinco tipos: pensamiento numérico, pensamiento espacial, pensamiento métrico, pensamiento variacional y pensamiento aleatorio y sistemas de datos. Es

este último el que correspondería a lo estadístico. En los estándares, se relaciona explícitamente cada uno de estos cinco tipos de pensamiento con subáreas de las matemáticas. Se dice que estos tipos de pensamiento no actúan de manera disyunta, sino que se superponen y todos deben desarrollarse conjuntamente. Por ejemplo, aunque el concepto de variabilidad puede ser transversal a dos tipos de pensamiento matemático, aleatorio y variacional, el tratamiento que se le da difiere en la medida en que en el primero se trata de dar sentido e interpretar y en el segundo se trata de modelarla usando un lenguaje formal.

En los estándares se define el pensamiento estadístico como:

> [aquel que]…ayuda a tomar decisiones en situaciones de incertidumbre, de azar, de riesgo o de ambigüedad por falta de información confiable, en las que no es posible predecir con seguridad lo que va a pasar…Ayuda a buscar soluciones razonables a problemas en los que no hay una solución clara y segura, abordándolos con un espíritu de exploración y de investigación mediante la construcción de modelos de fenómenos físicos, sociales o de juegos de azar… (pp. 64-65).

En la Tabla 3 se muestran las habilidades que deben dominar los estudiantes al finalizar el grado tercero de primaria.

Tabla 3

Estándares básicos de competencias para el pensamiento estadístico (o pensamiento aleatorio y sistemas de datos).

Primero a tercero
1. Clasifico y organizo datos de acuerdo con cualidades y atributos y los presento en tablas.
2. Interpreto cualitativamente datos referido a situaciones del entorno escolar.
3. Describo situaciones o eventos a partir de un conjunto de datos.
4. Represento datos relativos a mi entorno usando objetos concretos, pictogramas y diagramas de barras.
5. Identifico regularidades y tendencias en un conjunto de datos.
6. Explico –desde mi experiencia– la posibilidad o imposibilidad de ocurrencia de eventos cotidianos.
7. Predigo si la posibilidad de ocurrencia de un evento es mayor que la de otro.
8. Resuelvo y formulo preguntas que requieran para su solución coleccionar y analizar datos del entorno próximo.

Si bien la definición de los estándares no se ajusta exactamente a ninguna de las definiciones de pensamiento estadístico en la literatura, contiene elementos que son comunes a las definiciones entregadas por varios autores. Por ejemplo, si un estudiante domina las habilidades listadas en los estándares (Tabla 3) al finalizar grado tercero podría considerarse que se encuentra entre los niveles Informal e Inconsistente en el desarrollo del pensamiento estadístico según la propuesta de Watson y Callingham (2005) o podría clasificarse como un estudiante que se encuentra en estadios tempranos de alfabetismo estadístico (*Statistical Literacy).* En la Tabla 4 se establece la correspondencia entre las habilidades definidas en los estándares y las propuestas de Watson y Callingham (2005) y de Ben-Zvi y Garfield (2004).

Tabla 4

Correspondencia entre los estándares del MEN y las propuestas de Callingham y Watson y Ben-Zvi y Garfield.

Similitudes con los estándares de competencias del MEN (2006)	
Callingham y Watson (2005)	Relación idiosincrática con el contexto, habilidades básicas relacionadas con el conteo, explicaciones tautológicas. También se comparten elementos con el nivel informal, donde los estudiantes pueden realizar operaciones o comparaciones de un solo paso a un nivel informal. Además, se relaciona con el nivel inconsistente en la medida en que existen habilidades relacionadas con el uso preferente de interpretaciones cualitativas sobre los datos.
Ben-Zvi y Garfield (2004)	Corresponde con el alfabetismo estadístico. Conocimiento estadístico a un nivel informal. Nociones básicas de probabilidad y variabilidad. Para su evaluación se usan términos como identificar, describir, leer e interpretar. No requiere realizar cálculos, solamente conocimientos o nociones sobre los conceptos estadísticos.

En este trabajo se llamará pensamiento estadístico a lo que evalúan los EBC respecto al pensamiento aleatorio y sistemas de datos. Considerando las jerarquías establecidas por Ben-Zvi y Garfield (2004) y por Watson y Callingham (2003, 2005), se considera que el Pensamiento Estadístico posee una organización jerárquica en donde las personas pasan por varios estadios de dominio de los conceptos y habilidades estadísticas. En un caso se llaman niveles y en otro se conceptualizan como una triarquia.

Teniendo en cuenta el rendimiento en las pruebas como PISA y a las dificultades documentadas sobre el desempeño de estudiantes de diferentes niveles en el pensamiento estadístico, es de esperar que los estudiantes no dominen adecuadamente las habilidades planteadas en los estándares. La evaluación educativa debe ser capaz de identificar los perfiles de dominio de los estudiantes para poder obtener información que sirva para poder mejorar la instrucción de los estudiantes en el área de la estadística. Las Evaluaciones de Diagnóstico

Cognitivo pueden dar respuesta a esta necesidad ya que son capaces de retroalimentar de manera detallada el desempeño de los estudiantes en pruebas de aplicación masiva, superando su función sumativa.

Evaluación de Diagnóstico Cognitivo y Modelos de Diagnóstico Cognitivo

A finales de la década de 1980, todo el sistema educativo a nivel mundial demandaba que las evaluaciones tradicionales entregaran información útil para mejorar los procesos de aprendizaje de los estudiantes en todas las áreas. Esto derivó en la creación de evaluaciones que miden estructuras de conocimiento y procesos cognitivos específicos, éstas se conocen como Evaluaciones de Diagnóstico Cognitivo (Leighton y Gierl, 2007). En general, este tipo de evaluaciones se caracterizan por tener un fuerte sustento teórico en relación con los procesos cognitivos de las personas y además por utilizar modelos estadísticos robustos para modelar la relación entre las características de los ítems y los procesos de pensamiento de las personas, todo esto con un resultado informativo.

Cuando se implementan las Evaluaciones de Diagnóstico Cognitivo (en adelante, también llamadas EDC) la comunidad educativa tiene información detallada sobre el desempeño de los estudiantes. Las EDC conceptualizan lo que se mide como atributos multidimensionales con categorías discretas, de manera que el desempeño del estudiante en cualquier test ya no se describe como una proporción de aciertos o como su ubicación en un continuo de habilidad, sino que se puede precisar cuál es su patrón de dominio (o probabilidad de dominio) en términos de procesos cognitivos.

La función diagnóstica de las EDC ofrece una gran posibilidad para el sistema educativo en la medida en que la información que ofrecen puede utilizarse para entregar información personalizada a estudiantes, muy útil para los docentes y para quienes participan en la elaboración de políticas educativas. De hecho, puede decirse que las EDC trascienden la

diferenciación entre evaluación formativa y sumativa, en la medida en que puede cumplir ambas funciones y pueden utilizarse en cualquier momento dentro del proceso educativo.

En un trabajo clásico, Nichols (1994) planteó cinco pasos que se deben tener en cuenta a la hora de diseñar EDC:

1. Construcción teórica: Aquí se adopta o desarrolla el modelo y la estructura de conocimiento que se requiere para enfrentarse a la evaluación.
2. Selección del diseño: Al final de esta fase se busca diseñar ítems que correspondan con el modelo cognitivo especificado. Se busca que de acuerdo con el modelo del paso 1, hallan hipótesis de la manera en que los evaluados responderán los ítems.
3. Aplicación del test: De acuerdo con lo previamente definido se debe definir la forma en que se aplicará el test, cuál será el modo de aplicación.
4. Asignación del puntaje: Asignar puntuaciones que sean informativas sobre el desempeño del evaluado. Generar información útil sobre los procesos envueltos en sus respuestas.
5. Revisión del diseño: Se revisa el diseño de evaluación para ver la correspondencia entre el modelo cognitivo y éste. Se debe acumular evidencia respeto a la pertinencia y calidad de la evaluación.

Por su parte, DiBello, Roussoss & Stout (2007) mencionan seis pasos en el proceso de implementación de las EDC:

1. Descripción del objetivo de la evaluación: El objetivo de la evaluación debe ser explícito, lo que tener implicaciones en la descripción y pertinencia de lo que se precisa evaluar.

2. Descripción del modelo teórico: Teniendo en cuenta el punto anterior, se deben definir en detalle los atributos por evaluar, considerando la teoría adoptada.
3. Desarrollo de las tareas: La selección de las tareas de evaluación debe hacerse, preferiblemente, con base en lo que la evidencia haya demostrado que se ajusta mejor al objetivo y al contenido por evaluar.
4. Especificación del modelo psicométrico que relacione el desempeño con atributos: El modelo psicométrico escogido debe ser consistente con el objetivo de la evaluación, el modelo teórico adoptado y las tareas de evaluación.
5. Selección de los métodos estadísticos para estimación del modelo y evaluación de resultados: Esto hace referencia a la elección del modelo estadístico para la calibración de los indicadores de dominio de los evaluados.
6. Desarrollo del sistema para reportar los resultados a los estudiantes, profesores y otros: Este es un paso importante debido a que sirve para retroalimentar el desempeño. Además, la información aportada debe ser útil para los destinatarios de los informes.

Una diferencia importante entre estas dos propuestas es que en el momento en el que Nichols (1994) realizó sus planteamientos existían pocas alternativas respecto a los modelos psicométricos disponibles. En la actualidad, se llama Modelos de Diagnóstico Cognitivo (en adelante, MDC) a estos modelos psicométricos y estadísticos que ayudan a formular la probabilidad de que una persona acierte determinado conjunto de ítems dada cierta configuración de atributos que puede poseer o no (de la Torre & Minchen, 2014). Dicha probabilidad es función de los atributos que domine la persona y de los que estén previamente definidos como requisitos para contestar, así como su configuración (e.g., interdependencia).

La elección del MDC está en función de diferentes elementos, entre ellos la naturaleza del constructo por medir (e.g., jerárquico), la relación entre los atributos implicados en los ítems (e.g. suficiencia para acertar), entre otros. Los MDC hacen parte del cuerpo teórico de la psicometría y con ellos se han realizado aplicaciones propias de este campo de conocimiento, por ejemplo, avances en tests adaptativos informatizados (Liu, Ying & Zhang, 2015), detección de Funcionamiento Diferencial de los Ítems (Li & Wang, 2015), aplicaciones en pruebas situacionales (García, Olea & de la Torre, 2014), entre otros.

En los últimos años se han propuesto MDC que varían de acuerdo con la estructura y organización de los atributos que se quieran evaluar, por ejemplo, si estos se organizan jerárquicamente existen modelos más adecuados que otros. Desde el momento de su aparición, los MDC no han dejado de tener desarrollos, alcanzando a la fecha una mayor sofisticación, existiendo modelos con aplicaciones prácticas específicas (DiBello, et. al. 2007; Rupp, Templin, & Henson, 2010).

Existen diferentes criterios para clasificar los MDC; uno de ellos es la teorización sobre cómo interactúan los atributos en el momento de responder un ítem: puede ser de manera compensatoria o no compensatoria. Los modelos que asumen que un ítem se puede contestar acertadamente si el examinado posee el atributo A_x *o* A_y son llamados compensatorios. Mientras tanto, los modelos que suponen que para tener éxito en un ítem se requieren los atributos A_x *y* A_y son llamados no compensatorios (para ver una revisión completa de los MDC véase DiBello, Roussoss & Stout (2007)). Dado que la mayoría de las modelos poseen características específicas que restringen su aplicación a situaciones concretas, también se han

propuesto otros generalizados que se pueden adaptar a distintas condiciones (e.g. de la Torre 2011; von Davier, 2008).

Entre la variedad de MDC existentes, hay uno que fue pionero (Stout, 2002), que ha mostrado bastantes aplicaciones y que se encuentra bien documentado. Se trata del método *Rule Space* formulado por Tatsuoka (1983, 1990, 1991, 1995, 2009). Es importante anotar que en trabajos anteriores en castellano sobre este modelo se han propuesto traducciones (por ejemplo, Artavia-Medrano y Larreamendy-Joerns proponen "Método de Representación del Espacio de Reglas"), pero aquí se trabajará con el nombre original *Rule Space.*

El método Rule Space

El método *Rule Space* fue una de las primeras propuestas en el campo de los MDC y ha sido utilizado para evaluar el dominio de los estudiantes en áreas como la arquitectura (Katz, Martinez, Sheehan & Tatsuoka, 1993), comprensión de lectura (Buck, Tatsuoka & Kostin, 1997; Svetina, Gorin & Tatsuoka, 2011) y matemáticas (Chen, Gorin, Thompson & Tatsuoka, 2006; Dogan & Tatsuoka, 2008). El *Rule Space* se ha usado también para realizar comparaciones internacionales en el desempeño de los estudiantes en matemáticas (Tatsuoka, Corter & Tatsuoka, 2004), para realizar estudios longitudinales de desempeño y logro en matemáticas (Dean, 2006) y para evaluar el efecto del lenguaje en la validez de la prueba SAT (Examen de aptitudes para el acceso a la universidad en EE.UU.; Guerrero, 2001), entre otras. La mayoría de las aplicaciones del modelo se han manejado mediante la modalidad del *retrofitting* (en adelante también "ajuste retrospectivo"), que consiste en categorizar ítems de pruebas existentes según un conjunto de atributos definidos por expertos en el dominio de conocimiento evaluado y expertos en psicología cognitiva.

Evaluación de Pensamiento Estadístico con *Rule Space*

Trabajos como el de Artavia-Medrano y Larreamendy-Joerns (2012) y Gierl, Leighton y Hunka (2000) exponen en detalle en qué consiste este modelo; la siguiente explicación está basada en estos trabajos.

Desde *Rule Space*, el desempeño de una persona en un ítem o en un conjunto de ellos puede explicarse mediante la construcción de un perfil basado en el dominio o no-dominio de habilidades específicas. Estas habilidades específicas, que pueden ser conocimientos o procesos cognitivos, son llamadas *atributos* (Tatsuoka, 2009). Mientras que el perfil específico de dominio es llamado *Estado de Conocimiento* (en adelante también denotado como EC). El modelo asume que para responder de manera acertada a un ítem se requiere el dominio de todos los atributos implicados en él, es decir que se trata de un modelo no compensatorio (para mayor información sobre modelos compensatorios y no compensatorios ver Stout (2002)).

Rule Space debe su nombre a que en sus inicios tomó como modelo teórico la Teoría del Procesamiento de Información, en donde se asume que los ítems son problemas: para responder acertadamente, los estudiantes deben pasar de una situación inicial a una situación final teniendo en cuenta el *ambiente de la tarea* y *el espacio del problema;* aplicando *acciones y estrategias* que juntas pueden configurarse en *reglas de producción* (Artavia-Medrano & Larreamendy-Joerns, 2012). Las primeras aplicaciones del *Rule Space* trataron de identificar las reglas erróneas para darle sentido educativo y dirigir el proceso de enseñanza. Sin embargo, en algunos estudios se encontró que las reglas erróneas eran bastante inestables, mientras que los estudiantes con buen desempeño en las pruebas mostraban reglas de acción consistentes (e.g., Tatsuoka & Tatsuoka, 2005). Debido a esto, el enfoque del modelo pasó a la

clasificación de los estudiantes de acuerdo a los atributos necesarios para acertar (dominio y no-dominio), tomando como unidad de análisis las fuentes de los errores o componentes cognitivos relevantes. Las Teorías de Procesamiento de Información fueron utilizadas para realizar interpretaciones con sentido educativo y con una base teórica firme.

Como en cualquier EDC, la especificación de los atributos necesarios para resolver problemas debe contar con una base teórica y estar sustentada en evidencias de cómo las personas se enfrentan errónea o acertadamente a las tareas que se les presentan (por ejemplo, a partir de protocolos verbales); la información teórica se contrasta con el patrón de respuesta real y se clasifica a las personas en EC alrededor de ajustes al *patrón de respuesta ideal*. Esto último consiste en patrones de respuesta esperados de acuerdo a la correspondencia entre la configuración de atributos definida para cada ítem y la respuesta de los estudiantes a estos.

La clasificación estadística de las personas se realiza mediante el uso de dos parámetros: Theta (θ) que corresponde al parámetro de habilidad de las personas en un modelo de Teoría de Respuesta al Ítem y el parámetro Zeta (ζ) que es un índice de respuestas inusuales basado también en un modelo de Respuesta al Ítem (Cui, Gierl & Guo, 2017). Poner estos dos parámetros en un espacio bidimensional cartesiano es lo que se llama el *Rule Space* o el espacio de reglas.

El método *Rule Space* consta de dos fases. En la etapa de *identificación* se establece el conjunto de atributos por evaluar, se especifica su estructura y se prepara la construcción de las tareas, en el caso de estudios de ajuste retrospectivo se realiza un inventario de los atributos que requiere cada ítem para ser resuelto. En la segunda etapa, de *clasificación*, los

evaluados son asignados a estados de conocimiento, según los análisis estadísticos que utiliza *Rule Space*. Sin embargo, en medio de estas dos etapas, existe una tarea importante que es el desarrollo de las tareas de evaluación y su respectiva implementación.

Etapa de identificación.

Para comenzar con esta etapa se debe definir la lista de atributos por evaluar y en caso de existir una relación entre ellos, debe especificarse. Esto debe realizarse, con preferencia, considerando los desarrollos teóricos más recientes sobre qué y cómo aprenden los estudiantes en el área de conocimiento en cuestión. La Figura 2 muestra un ejemplo basado en una estructura hipotética de 6 atributos.

Dada la lista inicial de atributos y establecida su organización, se procede a crear una matriz A o matriz de adyacencia, de forma *j* x *k*, en donde se hace explícita la relación específica entre ellos. En la Figura 3 se muestra la matriz A que corresponde a la estructura propuesta para estos atributos; en esta matriz se codifica como 1 cuando un atributo está subordinado a otro, o en otras palabras cuando depende de otro. En la primera fila de la matriz se indica que el atributo A_1 es prerrequisito directo de los atributos A_2 y A_3, asimismo se observa que el atributo A_2 es prerrequisito del atributo A_4, que el atributo A_3 es prerrequisito del A_6 y que el A_4 es prerrequisito del A_5. Cuando hay filas en ceros quiere decir que ese atributo no es prerrequisito directo de ningún otro, como en el caso de A_5 y A_6.

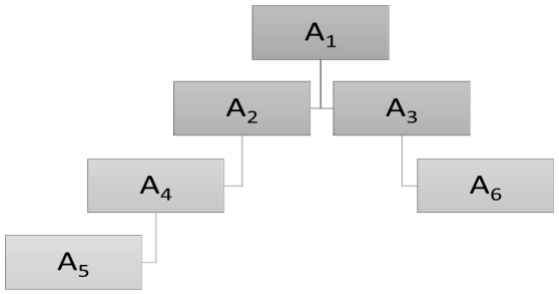

Figura 2. Estructura jerárquica de 6 atributos.

Mediante una transformación booleana, la matriz A se convierte en una matriz R que puntualiza la dependencia directa e indirecta entre los atributos. Por ejemplo, se especifica la relación de dependencia indirecta de A_5 frente a A_1. Esta es la matriz de *accesibilidad,* denotada como matriz R (Figura 4). Siguiendo la configuración de atributos hipotética que se ha propuesto, en la primera fila indica que A_1 es prerrequisito de sí mismo y de los demás atributos; asimismo, al observar la segunda fila se infiere que A_2 es prerrequisito de sí mismo y de A_4. Por último, A_5 y A_6 solamente son prerrequisito de ellos mismos.

En seguida, se establece la matriz Q (de forma k x p), que es el conjunto potencial de ítems dado cierto número de atributos. La expresión para calcular el número potencial de ítems es 2^k-1, en este caso es 2^6-1; es decir 63 ítems potenciales que se muestran en el Anexo 1.

$$A = \begin{matrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Figura 3. Ejemplo de matriz de adyacencia.

$$R = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Figura 4. Ejemplo de matriz de accesibilidad.

Dado que a partir de la matriz *R* se sabe que hay ítems que deben cumplir con las características que impone la relación jerárquica entre los atributos, se descartan aquellos que no las cumplen. Por ejemplo, no se tiene en consideración aquellos en donde no se requiere el atributo 1 (es decir, que tienen un valor de 0 en la primera fila) debido a que éste es prerrequisito de los demás y de él mismo. Debido a esto, es necesario tener en cuenta todas las restricciones y se construye la *matriz Q reducida* (Q_r; de forma *k* x *i*), que al igual que la matriz *Q* se obtiene a partir de operaciones booleanas. La matriz Q reducida se muestra en la Figura 5.

$$Qr = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{matrix}$$

Figura 5. Ejemplo de matriz de incidencia reducida.

La matriz R y la matriz Q se complementan dado que la primera especifica la jerarquía de los atributos como punto de partida, y la segunda aporta información sobre las características de los ítems que se deben construir como materialización de los patrones de respuesta esperados.

La matriz Q_r es la guía que deben tomar los constructores de ítems en el caso de desarrollo de instrumentos, o los expertos que estén adaptando el modelo en caso de la aplicación a tests ya existentes, respecto a las combinaciones de atributos permitidas.

Las pruebas deben ser desarrolladas con base en lo anterior, pero también se debe considerar cuestiones como la representatividad de todos los atributos en la prueba y asegurar que las tareas cumplan con los requisitos que impone cada uno de los atributos definidos. En este sentido, pueden utilizarse los formatos de ítems que sean necesarios para obtener

evidencias sobre los atributos evaluados y además se debe considerar lo que la teoría adoptada para el desarrollo de la prueba indique.

Etapa de clasificación.

Una vez aplicada la prueba, se procede a estimar el parámetro de habilidad (θ) generalmente utilizando el modelo de Rasch, y el de *atipicalidad* (ζ); se representan en un plano con coordenadas (θ,ζ) y se mide la distancia entre el patrón de respuesta observado y el patrón de respuesta ideal. De esta manera, se puede identificar posibles grupos de clasificación y así extraer información sobre procesos cognitivos o caminos de acción típicos (no necesariamente que conduzcan a aciertos). En el parámetro de *atipicalidad* se contempla el desajuste entre el patrón de respuestas observadas y las esperadas de acuerdo al modelo, como, por ejemplo, fallar un ítem "fácil" o acertar un ítem "difícil" dado cierto nivel de habilidad. Este insumo es importante para trabajar sobre las posibles agrupaciones de patrones de respuesta erróneos en los estudiantes e implementar estrategias de mejora o identificar Estados de Conocimiento que no se contemplaron antes por parte de los diseñadores de la prueba o de los constructores de la lista de atributos.

Con la información que arroja *Rule Space* es posible retroalimentar de manera eficiente el proceso de aprendizaje de los estudiantes, objetivo principal de cualquier Evaluación de Diagnóstico Cognitivo.

El objetivo principal de este trabajo será desarrollar una Evaluación de Diagnóstico Cognitivo, utilizando el método *Rule Space* para evaluar el pensamiento estadístico en niños

del primer ciclo de básica primaria en Colombia. La selección de los niños de primer a tercer grado (primer ciclo) de primaria obedece a la importancia que tiene identificar en detalle el desempeño de los estudiantes a este nivel y así proponer estrategias para mejorar su proceso de aprendizaje en el área de la estadística.

El objetivo principal de este trabajo es llegar a un diagnóstico detallado sobre el desempeño de los estudiantes de primaria en el dominio de pensamiento estadístico. Para lograrlo, se definieron dos objetivos específicos:

1. Desarrollar un instrumento de evaluación del Pensamiento Estadístico para estudiantes de primero a tercer grado de básica primaria.
2. Implementar análisis propios con el método *Rule Space*.

Método

Para realizar la evaluación diagnóstica propuesta fue necesario dividir el método en dos fases: la primera consiste en el desarrollo de un instrumento de evaluación del Pensamiento Estadístico, mientras que la segunda se ocupa de la aplicación del método *Rule Space*.

Fase 1 – Desarrollo del Instrumento

Esta fase corresponde al objetivo específico del desarrollo de un instrumento de evaluación del pensamiento estadístico en niños de primer ciclo de básica primaria, tomando como base para su construcción los lineamientos curriculares que dan los EBC del Ministerio de Educación Nacional.

Participantes.

Se realizó una convocatoria a docentes del área de matemáticas para participar como expertos en la etapa de definición de componentes, en el diseño de las especificaciones de la prueba y en la validación de los ítems. Los criterios de inclusión para los profesionales fueron (a) estar vinculado a una institución educativa oficial o no oficial como docente, (b) tener experiencia en la enseñanza de la matemática en el nivel de básica primaria de mínimo cinco años y (c) conocer la conceptualización de pensamiento matemático contenida en los EBC. La información para participar fue difundida entre docentes de diferentes instituciones educativas de la ciudad de Bogotá y los seleccionados fueron aquellos que cumplieron con los criterios de inclusión y que manifestaron contar con disponibilidad. Después de esto, se conformó un equipo de cuatro expertos, sus perfiles se muestran en la Tabla 5.

Tabla 5

Perfil de los profesionales seleccionados para participar como expertos en el estudio.

Experto	Perfil
Experta 1	Títulos Universitarios: Psicóloga, Especialista en Gerencia Educativa. Experiencia: Docente del sector oficial desde 1982. Experiencia específica de 15 años en docencia de las matemáticas en primaria.
Experto 2	Título Universitario: Psicólogo, Especialista Experiencia: 17 años de experiencia como docente en el área de matemáticas, primero en enseñanza en primaria durante cinco años, al momento de ser seleccionado, impartía en los primeros grados de bachillerato.
Experta 3	Título Universitario: Licenciada en Matemáticas. Experiencia: Cerca de nueve años de experiencia en enseñanza de matemáticas en primaria. Siempre ha trabajado en un colegio femenino del sector no oficial.
Experto 4	Título Universitario: Ingeniero de Sistemas Experiencia: ocho años en enseñanza de matemáticas en instituciones del sector oficial, en primaria.

Los expertos acompañaron toda la fase de construcción del instrumento de pensamiento estadístico, excepto el experto 4 quien asistió solo a la primera sesión.

Para esta fase del estudio también se contó con la participación de 221 estudiantes de un colegio No Oficial ubicado en el norte de la ciudad de Bogotá, a quienes les fue aplicada la versión preliminar de la prueba de pensamiento estadístico. Los estudiantes eran niños y niñas de los grados segundo (71), tercero (76) y cuarto (74).

Instrumentos.

Para el desarrollo de la prueba se diseñaron diferentes instrumentos que sirvieron para organizar la información, pero también como guía de diferentes actividades. Se tomaron

modelos diseñados para investigaciones realizadas en el Laboratorio de Psicometría de la Universidad Nacional de Colombia y se adaptaron de acuerdo con las necesidades particulares de esta investigación.

Con el objetivo de estructurar las actividades por desarrollar con los docentes, se construyó un protocolo de panel de expertos (ver Anexo 2). Esta herramienta sirvió para asegurar que el objetivo del panel se cumpliera de acuerdo con lo planeado. El protocolo estructuró en cuatro partes el plan de trabajo: la primera consistió en realizar un repaso detallado sobre los conceptos propuestos en los EBC sobre el Pensamiento Matemático y especialmente aquello que corresponde al *Pensamiento Estadístico*, la segunda actividad estuvo encaminada a definir la lista de componentes por evaluar, la tercera a la construcción de las especificaciones de la prueba y la última a validar los ítems.

Para facilitar la organización y el examen de la información respecto a la lista de componentes, se diseñó un formato de validación de esta (ver Anexo 3). Este formato buscó simplificar la labor de los docentes al momento de consignar sus observaciones sobre cada componente. En este punto se llama componentes a los elementos que se identificarían como importantes para evaluar. No se le llamaron atributos debido a que se tomará como base lo contenido en los EBC, es decir, la lista será orientada de acuerdo con el currículo y no a una teoría cognitiva.

Con el fin de asegurar que el desarrollo de los ítems fuera consistente con la información recabada en las primeras etapas del panel de expertos, se construyó un protocolo detallado para la construcción de ítems, que se muestra en el Anexo 4. Allí se incluye una corta

introducción sobre el dominio por evaluar, además se exponen los componentes definidos previamente, se detallan las necesidades de construcción de acuerdo con las especificaciones de la prueba y se hacen recomendaciones sobre elementos técnicos que se deben considerar.

Para asegurar que los ítems fueran elaborados de manera estructurada y el constructor proporcionara toda la información sobre cada uno de ellos, se desarrolló un formato de construcción de ítems. Este también sirvió para presentar los ítems para la validación. Teniendo en cuenta el tipo de ítems por construir, se estableció un formato para ítems de selección múltiple y otro para los de respuesta construida. Estos formatos se muestran en el Anexo 5.

En vista de que los ítems de respuesta construida requieren del desarrollo de criterios de calificación claros, fue necesaria la elaboración de una guía en donde se establecieron dichos criterios para estos ítems. Esta guía corresponde al protocolo de calificación que se encuentra en el Anexo 6.

También fue elaborado un formato de consentimiento informado que fue entregado a cada estudiante con anterioridad al pilotaje de la prueba. Esto debido a que los participantes son menores de edad y se requería la autorización de los padres de familia para hacer parte del estudio. En el consentimiento se informaba a los padres de familia los objetivos generales del estudio, se especificaba que no se utilizarían datos sensibles y que la información recogida sería tratada siguiendo los principios de confidencialidad y no maleficencia. Este formato se encuentra en el Anexo 7.

Procedimiento.

Se partió del hecho de que, aunque se buscaba crear un instrumento para implementar el *Rule Space,* este modelo y en general las EDC tienen como requisito elemental que la parte teórica que da sustento a la prueba esté fundamentada adecuadamente. Sin embargo, esto no fue así en este caso debido a que los EBC son lineamientos curriculares y no corresponden a un planteamiento sustentado empírica o teóricamente. Por este motivo, se decidió llamarles componentes y no atributos, para evidenciar esta situación.

Para desarrollar la prueba de pensamiento estadístico el primer paso fue la validación de la lista de componentes, el segundo paso correspondió a la definición de especificaciones de la prueba y la elaboración y validación de los ítems, y en tercer lugar se desarrolló un pilotaje con la versión inicial de la prueba para realizar ajustes finales.

Elaboración de la lista de componentes.

Para determinar los componentes de pensamiento estadístico que se utilizarían como base para el diseño del instrumento se desarrolló un procedimiento de panel de expertos cuya estructura se consignó en el protocolo descrito anteriormente y que sirvió como guía para desarrollar esta actividad. La elaboración de la lista de componentes tomó dos sesiones presenciales con una duración total aproximada de 16 horas.

Primero se buscó que los expertos demostraran su conocimiento sobre la estructura y conceptos básicos de los EBC para el área de matemáticas y específicamente para lo que allí

se llama *pensamiento aleatorio y sistemas de datos*. En este punto se realizó un repaso de los EBC y se promovió la discusión alrededor de sus características y su contenido.

Después del examen de los EBC dentro del panel, se sometió a consideración por parte de los expertos una propuesta inicial de la definición de pensamiento estadístico y una lista preliminar de componentes. La definición de pensamiento estadístico inicial correspondió a la de *pensamiento aleatorio y sistemas de datos* tal como aparece en los EBC. Por su parte, la lista preliminar de componentes se basó en las especificaciones que los estándares entregan sobre las actividades que todo estudiante debería ser capaz de dominar "al terminar grado tercero", para el *pensamiento aleatorio y sistemas de datos*. También se realizó una revisión de los Derechos Básicos de Aprendizaje en Matemáticas (MEN, 2006) para primer ciclo. La definición de pensamiento estadístico inicial fue:

> *El Pensamiento Estadístico ayuda a buscar soluciones razonables a problemas en los que no hay una solución clara y segura, abordándolos con un espíritu de exploración y de investigación mediante la construcción de modelos de fenómenos físicos, sociales o de juegos de azar y la utilización de estrategias como la exploración de sistemas de datos, la simulación de experimentos y la realización de conteos.* Tomado de MEN (2006), pp 66-67.

Por su parte, la lista inicial de componentes consistió en 17 en total: 5 componentes generales y 12 específicos. Estos se muestran en la Tabla 6.

Tabla 6

Propuesta inicial de la lista de componentes del Pensamiento Estadístico.

Componentes Generales	
1 - Generación, interpretación y transformación de gráficos	
2 - Estimación y aproximación	
3 - Comunicación, representación y modelación	
4 - Razonamiento y argumentación	
5 - Planteamiento y resolución de problemas	
Componentes Específicos	
1 - Clasificar y ordenar datos	7 - Interpretar lo que un diagrama de barras determinado representa
2 - Identificar regularidades y tendencias de un conjunto de datos	8 - Interpretar lo que una tabla de frecuencias representa
3 - Describir características (patrones) de un conjunto (situaciones) a partir de los datos que lo representan	9 - Describir tendencias a partir de los datos
4 - Representar un conjunto de datos a partir de un diagrama de barras	10 - Establecer conjeturas sobre la posibilidad de ocurrencia de ciertos eventos
5 - Representar un conjunto de datos a partir de una tabla de frecuencias	11 - Resolver problemas a partir de datos recolectados
6 - Transformar los datos de una representación a otra	12 - Resolver problemas a partir de la estimación de grados de posibilidad de ocurrencia

Esta propuesta fue discutida en el panel durante la primera sesión de este, a partir de lo cual se propusieron modificaciones y al final el grupo adoptó una definición clara del dominio, del tipo de conocimiento y las facetas por evaluar (terminología de los EBC).

En esta misma sesión, los expertos realizaron una discusión sobre la pertinencia de cada uno de los componentes propuestos y propusieron modificaciones. En primer lugar, se decidió

eliminar los componentes generales por considerarlos transversales a varios tipos de conocimiento o pensamiento (e.g., "razonamiento y argumentación") y difíciles de aislar a la hora de evaluarlos; respecto a los específicos se aclararon algunos (e.g., "identificar regularidades" se cambió por "identificar patrones") y se incluyeron otros (e.g., "interpretar lo que una tabla de frecuencias representa"). En la tabla 7 se muestra esta nueva lista que consta de ocho componentes por evaluar.

Esta lista fue revisada después de la primera sesión del panel, de manera que se pudieran presentar observaciones y recomendaciones de modificación, así como preguntas para promover el debate durante la segunda sesión. Así, se plantearon cuestiones principalmente referidas a criterios de exhaustividad de la lista de componentes (e.g., ¿Falta algún componente que se enseñe en el aula y no esté incluido en la lista? ¿Los componentes definidos abarcan todo el dominio de lo que se ha definido como pensamiento estadístico?), la relación entre estos (e.g., ¿Para dominar el componente *x*, es necesario dominar el componente *y*?) y la proyección sobre la evaluación de estos componentes en una prueba escrita (¿Cómo se evaluaría *z*?).

Tabla 7

Lista de componentes modificada.

Número	Habilidades que el estudiante debe dominar
1	Clasifico y organizo datos de acuerdo con cualidades y atributos y los presento en tablas.
2	Interpreto cualitativamente datos referidos a situaciones del entorno escolar.
3	Describo situaciones o eventos a partir de un conjunto de datos.
4	Represento datos relativos a mi entorno usando objetos concretos, pictogramas y diagramas de barras.
5	Identifico regularidades y tendencias en un conjunto de datos.
6	Explico –desde mi experiencia– la posibilidad o imposibilidad de ocurrencia de eventos cotidianos.
7	Predigo si la posibilidad de ocurrencia de un evento es mayor que la de otro.
8	Resuelvo y formulo preguntas que requieran para su solución coleccionar y analizar datos del entorno próximo.

Después de obtener esta lista, se realizó una nueva ronda de revisión, presentando las mismas cuestiones para la discusión. Con esto, los expertos realizaron las modificaciones finales que principalmente obedecieron a la necesidad de aclarar varios de los componentes definidos. Específicamente, se separaron los componentes que antes estaban juntos en "clasificar y ordenar datos" y en "identificar patrones y tendencias", se incluyó la descripción de características comunes de conjunto de datos y la representación e interpretación de pictogramas y se eliminaron los componentes relacionados con "resolver problemas" por considerarlos confusos.

Una vez definida la lista de atributos se verificó si los componentes eran coherentes con la definición de pensamiento estadístico y con la faceta por evaluar. Finalmente se pidió a los expertos que elaboraran la definición de cada uno de los componentes.

Diseño de especificaciones de la Prueba de Pensamiento Estadístico.

Una vez establecida la lista de componentes con sus definiciones, se procedió a la construcción de las especificaciones de la prueba. Para esto sirvieron como guía los estándares de desarrollo de instrumentos de AERA, APA y NCME (2014). La actividad consistió en promover la discusión dentro del panel de expertos sobre aspectos clave para la elaboración de los ítems como el propósito de la prueba, el formato de los ítems, población objetivo, aspectos culturales, forma de calificación, condiciones materiales de aplicación, entre otras. Este trabajo se realizó al final de la segunda sesión de panel de expertos, con una duración aproximada de 2 horas.

Como apoyo adicional para la construcción de ítems, los expertos también trabajaron en la propuesta de una estructura de prueba. Esta estructura se diseñó para especificar la combinación de componentes en cada ítem y el formato por usar y así facilitar la labor de los constructores de ítems. La estructura se anexó al protocolo de construcción de ítems.

Elaboración y validación de ítems.

La construcción de ítems estuvo a cargo de un profesional con experiencia en la elaboración de pruebas en el contexto educativo, apoyado por el protocolo de construcción y los formatos de desarrollo de ítems. El desarrollo de los ítems fue una tarea independiente del

panel de expertos y se realizó en un periodo de aproximadamente dos semanas usando las especificaciones para la prueba, la estructura de prueba, los formatos de construcción y los estándares para el desarrollo de pruebas de AERA, APA y NCME (2014). Una vez finalizada la labor de construcción de ítems, fueron preparados para presentarse en el panel de expertos para su respectiva validación.

El proceso de validación se desarrolló en dos sesiones: La primera fue presencial y se llevó a cabo dos semanas después de la reunión en la que se elaboraron las especificaciones de la prueba y se concertó la estructura de esta. La segunda se realizó una semana después de la primera y fue necesario utilizar la modalidad de teleconferencia debido a que uno de los expertos no podía asistir al lugar habilitado para las reuniones.

Sesión 1. Se presentó la versión inicial de 20 ítems, de los cuales se aprobaron 15 (algunos de ellos con modificaciones realizadas por el equipo completo en el transcurso de la sesión), mientras que 4 de ellos recibieron sugerencias de modificación, del restante se recibió la sugerencia de eliminación.

Sesión 2. Fueron aceptados aquellos ítems sobre los que se solicitaron cambios o ajustes (4 ítems) en la sesión anterior y se presentaron 4 nuevos ítems (los faltantes según la estructura propuesta) que fueron aceptados con modificaciones menores. Después se realizó una verificación de que los ítems evaluaran los componentes propuestos para cada uno de ellos mediante el enmascaramiento de los componentes y el ejercicio de codificación retrospectiva, presentándose una coincidencia adecuada.

Las imágenes de apoyo utilizadas en toda la prueba fueron tomadas de *Pixabay*[1] y *Freepic*[2] que son bancos de imágenes gratuitos y con derechos de autor liberados.

Una vez aprobados todos los ítems que conformarían la prueba, se procedió a organizarlos de acuerdo con lo establecido por los expertos en las sesiones de validación. Los expertos también sugirieron que se incluyera una página de preámbulo en donde además de consignar algunos datos básicos, los estudiantes se encontraran con una explicación breve de cada una de las tres formas de representación gráfica que evaluaría la prueba y así evitar posibles errores por el desconocimiento de sus denominaciones. En esa página también se indicó el nombre de la prueba, la afiliación institucional de la investigación y un agradecimiento por participar.

Esta página inicial más los ítems se pusieron juntos formando la versión preliminar de la prueba, que fue enviada por correo electrónico a cada uno de los expertos, los cuales la aprobaron. El programa utilizado para la diagramación de la prueba fue Microsoft Word; la fuente seleccionada fue *Times New Roman* en tamaño 12.

Pilotaje.

La actividad de aplicación piloto de la prueba sirvió para tomar los tiempos promedio que emplearon los estudiantes. Además, se recibieron comentarios por parte de profesores y

[1] https://pixabay.com/es/

[2] https://www.freepik.es/

alumnos sobre la diagramación de la prueba, sobre claridad de los ítems y se recibieron sugerencias sobre algunos cambios principalmente de forma para la prueba en general. Una vez recogida toda la información, se consultó con los expertos (vía correo electrónico) la pertinencia de dichas observaciones y se realizaron los cambios que ellos aprobaron.

Fase 2 – Implementación Rule Space

Participantes.

Inicialmente, la prueba fue aplicada a un total de 1580 estudiantes entre segundo y quinto grado de educación básica primaria de las ciudades de Bogotá, La Paz y Valledupar (Cesar), Tumaco (Nariño) y San José del Guaviare (Guaviare). La selección de los centros educativos se realizó por conveniencia. Los colegios eran invitados directamente a hacer parte del estudio y si accedían se entregaba toda la información sobre los propósitos y el alcance de la investigación.

La muestra fue reducida inicialmente a 1535 estudiantes debido a que se eliminaron aquellos que presentaron porcentaje de omisiones en la prueba por encima del 30%. Con esta muestra se realizaron los análisis psicométricos básicos de la Prueba de Pensamiento Estadístico.

Para realizar los análisis propios del método *Rule Space* fue necesario trabajar con una muestra de 562 estudiantes. Esto debido a que el software utilizado no admite omisiones y este grupo de estudiantes presentó una cadena de respuestas completa.

Instrumentos.

Para obtener la autorización para la participación de los estudiantes fue necesario utilizar el formato de consentimiento informado diseñado en la primera fase de esta investigación. Este documento fue entregado con anticipación a la aplicación de la prueba para que fuera devuelto por los estudiantes debidamente diligenciado y firmado por un acudiente reconocido por el respectivo colegio.

Para evaluar el pensamiento estadístico en los estudiantes que cumplieron con el requisito previo de la presentación del consentimiento informado se utilizó la Prueba de Pensamiento Estadístico resultado de la fase I que consta de 24 ítems, 8 de selección múltiple y 16 de respuesta construida.

Para realizar la calificación de las pruebas aplicadas se utilizó el protocolo de calificación diseñado en la fase anterior, que incluyó las modificaciones realizadas después del pilotaje de la versión preliminar de la prueba de pensamiento estadístico.

Procedimiento.

Aplicación del instrumento.

Para comenzar la tarea de recolección de datos se conformó un equipo de aplicación del que hicieron parte varios estudiantes de pregrado y posgrado del Departamento de Psicología de la Universidad Nacional de Colombia. Después, se invitó a varias instituciones educativas de diferentes ciudades colombianas para participar en el estudio. La convocatoria para los colegios se hizo mediante correo electrónico y una vez manifestaban su interés por participar,

una persona del equipo realizaba una visita preliminar en la que informaba a los directivos los pormenores de la investigación. También se hacía entrega del consentimiento informado para que los estudiantes los llevaran a sus acudientes y los devolvieran firmados antes de la aplicación. Esta visita también servía para recoger información previa para planear adecuadamente la visita de aplicación.

En la fecha y hora acordadas, el equipo encargado asistía a la respectiva institución educativa y realizaba la aplicación de la Prueba de Pensamiento Estadístico. Las condiciones de aplicación exigían que el lugar fuera el salón o uno de los salones en el que solían recibir clase los estudiantes y que el docente encargado en el momento permaneciera dentro del aula. Al finalizar la aplicación el equipo encargado procedió a rotular y a organizar el material para transferirlo al personal encargado de la calificación de las pruebas.

La calificación estuvo a cargo de dos psicólogos con experiencia en evaluación educativa, quienes siguieron la guía establecida para ello (Anexo 5). Para asegurar la unidad de criterio, se calificaron conjuntamente un grupo de 192 cuadernillos y se calculó el índice de acuerdo inter-jueces Kappa. Aquellos casos en los que se presentaron discrepancias sirvieron para ser revisados y así mejorar la unidad de criterio.

La calificación de todos los ítems fue dicotómica, asignando 1 en caso de acierto y 0 en caso de desacierto. Se consideraba como omisión cuando el estudiante no marcaba en ítems de selección múltiple o si en ítems de desarrollo el espacio destinado para la respuesta era dejado en blanco.

Análisis Psicométricos.

Los análisis psicométricos de la prueba empezaron por indicadores descriptivos respecto a las puntuaciones, así como un análisis de diferencias de medias en el desempeño en la prueba tomando como factores las variables de género, grado, ciudad, tipo de colegio y colegio.

Después, se realizó una revisión detallada de las características psicométricas de cada uno de los ítems utilizando indicadores de la Teoría Clásica de los Test y el modelo de Rasch. Con el fin de conservar los ítems con calidades psicométricas adecuadas para los análisis propios con *Rule Space*, se estableció que aquellos que reunieran dos o más de las siguientes características serían eliminados:

C1. Bajos índices de correlación ítem-prueba

C2. Índice de consistencia interna que aumenta al eliminar el ítem

C3. Desajuste cercano (infit)

C4. Desajuste lejano (outfit)

Una vez realizada esta depuración, se exploró la distribución de las puntuaciones entre los estudiantes. Por otra parte, se calcularon los índices de dificultad y de discriminación según la TCT utilizando el paquete de R llamado “psych”. También se corrió el modelo de Rasch utilizando los paquetes “eRm” y “ltm”, hallando el parámetro de dificultad de los ítems, la distribución de habilidad de los evaluados y las medidas de desajuste para los ítems y las personas.

Rule Space – Identificación de atributos

Para establecer los atributos que conformarían la matriz Q y así realizar los análisis de *Rule Space,* se comenzó por realizar una correspondencia directa entre la lista de componentes utilizada para el desarrollo de la prueba de pensamiento estadístico y los atributos. Por lo que la lista inicial de atributos fue de 14. Con esta información, se realizó un análisis de regresión múltiple con el fin de verificar si la configuración de atributos en los ítems (información contenida en la estructura de la prueba) podía explicar la varianza en la dificultad en estos. Al realizar este análisis se encontró que la configuración de los componentes tan solo explicaba el 30% de la varianza de la dificultad. Cuando esto ocurre, es necesario revisar la lista preliminar de atributos (e.g. Artavia Medrano, 2014).

Dicha revisión dio lugar a modificaciones en la lista de atributos. Estos cambios se realizaron con base en las siguientes consideraciones:

a) De los ítems en los que estaba implicado el componente "Clasificar datos", uno había sido eliminado (ítem 14) y en los demás se observó que implicaba más el conteo que la clasificación en la medida en que las categorías estaban explícitas en todos los ítems. Por estos motivos, este componente no fue tenido en cuenta como un atributo. Como reemplazo de este, se estableció como atributo "Determinar frecuencias dentro de un conjunto de datos".
b) El componente "Describir características comunes" fue modificado por "Identificar características comunes dentro de un conjunto de datos". Esto porque se observó que

los ítems realmente no exigían que el estudiante describiera características, solo identificarlas.

c) "Identificar tendencias" fue reemplazado por "Completar secuencias" dado que esto último fue lo que los ítems solicitaban al estudiante.

d) Los componentes de representación gráfica fueron unificados en uno solo, considerando que implican la transformación de datos de una representación a otra. El atributo que sintetiza a los tres componentes relacionados con gráficos es "Representar gráficamente información dada". Esta consideración también llevó a prescindir del componente "Encontrar la correspondencia entre dos formas de representación de datos".

e) Los componentes de interpretación de las tres formas de representación gráfica (barras, tabla y pictograma) fueron resumidos en uno solo llamado "Extraer información a partir de representaciones gráficas". Este nuevo nombre implica la especificación de la respuesta que elicitan los ítems.

f) Los componentes "Establecer conjeturas sobre la posibilidad de ocurrencia de ciertos eventos" y "Predecir si la posibilidad de ocurrencia de un evento es mayor que la de otro" fueron unidos en no solo debido a que ambos implican "Estimar la probabilidad de ocurrencia de ciertos eventos"

g) Adicionalmente, fueron incluidos los atributos de "Identificar el elemento mínimo y/o máximo de un conjunto" e "Interpretar situaciones descritas en el enunciado". El primero debido a que varios ítems solicitan que el estudiante haga esta tarea específica, que es diferente a contar y no puede considerarse como clasificar. El

segundo se planteó para no desconocer el hecho de que algunos problemas planteados en el cuerpo de ítems tienen cierto grado de complejidad y es necesario que la interpretación de ellos sea adecuada para que los estudiantes puedan enfrentarse de manera adecuada a la tarea presentada.

h) Los componentes "Ordenar datos", "Identificar el patrón dentro de un conjunto de datos" se mantuvieron.

Esto dio lugar a una nueva lista de atributos que implicó la realización de un procedimiento de codificación de los ítems con base en ella. Al final de este ejercicio se volvió a realizar el análisis de regresión múltiple para verificar si la configuración de atributos podía explicar la varianza en la dificultad de los ítems (estimada con el modelo de Rasch y con la TCT).

El resultado de las regresiones múltiples hace que el análisis con *Rule Space* sea viable en la medida en que la configuración de atributos es capaz de explicar la varianza en la dificultad de los ítems y posiblemente el desempeño de los evaluados en la prueba.

Rule Space – Construcción de matrices

Una vez codificados los ítems teniendo en cuenta la nueva lista de atributos, se realizó el ejercicio de elaboración de las matrices A y R: aunque por no existir evidencias respecto a la relación entre los atributos, la primera resulte en una matriz nula (todas las entradas igual a 0) y la segunda en una matriz de identidad (todas las entradas igual a 0 excepto en la diagonal). Además se elaboró la matriz Q a partir de los atributos definidos y a la respectiva codificación

de los ítems. La matriz Q fue utilizada como *input* para realizar los análisis propios del método *Rule Space.*

Antes de realizar los análisis relacionados con la clasificación de las personas se realizaron algunos análisis descriptivos respecto a los atributos y a los ítems. En relación con los atributos, se halló la cantidad de veces que se evalúa cada uno de ellos dentro de la prueba. Respecto a los ítems se encontró la cantidad de atributos y cuáles evalúa cada uno de ellos, y se relacionó con su parámetro de dificultad.

Rule Space - Clasificación de los evaluados

Las probabilidades de dominio de atributos que entrega el método *Rule Space* representan un valor diagnóstico importante; se pueden obtener y analizar de manera individual y colectiva. Se puede utilizar el promedio de dichas probabilidades para representar la información de dominio de atributos en términos generales, esto indicaría el desempeño grupal de los estudiantes en la prueba y en cada atributo. Por ejemplo, probabilidades promedio bajas indicarían que el conjunto de la muestra tiene una menor posibilidad de dominio, mientras que altas probabilidades promedio indicarían que la mayoría de evaluados tienen mayor posibilidad de dominar el atributo.

Por otra parte, con la información sobre las probabilidades de dominio para cada evaluado en cada uno de los atributos, se puede hallar el patrón de dominio de atributos para cada uno de ellos, que da lugar a la clasificación de los estudiantes en diferentes estados de conocimiento (en adelante, también llamados EC). Para cada patrón observado de respuestas, se calcula el valor de D^2 (distancia de Mahalanobis) entre dicho patrón y todos los patrones

ideales en sus cercanías. Para esto, se utilizó como criterio de clasificación que existiera una distancia menor a 4.5 ($D^2 < 4.5$) entre el patrón observado de respuestas de cada evaluado y el estado ideal de conocimiento más cercano.

Como parte de los análisis propios del *Rule Space* es posible reducir los EC, debido a las condiciones previas dadas por la matriz Q. Cuando hay un número elevado de estados de conocimiento, se debe realizar una agrupación de ellos y poder realizar análisis a nivel colectivo que tengan significado para poder mejorar el perfil de dominio de atributos de los estudiantes. Para ello, se realizó un análisis de conglomerados de *K* medias, el cual, a partir de las similitudes entre los estados de conocimiento definidos inicialmente, realiza agrupaciones hasta encontrar grupos homogéneos, considerando las probabilidades de dominio de atributos obtenido con el *Rule Space.* A dichas agrupaciones se les conoce con el nombre de Estados Conglomerados de Conocimiento (e.g., Artavia-Medrano, 2015).

Chen, Gorin, Thompson y Tatsuoka (2008) proponen los siguientes criterios para obtener un número apropiado de conglomerados: 1) la cantidad de estudiantes que se clasifiquen en cada conglomerado debe ser mayor que el 1% del tamaño de la muestra, 2) la distancia promedio entre cada estado de conocimiento y el centro del estado de conocimiento hipotético debe ser menor que 2, y 3) el valor del estadístico *F* como resultado del ANOVA de un solo factor para cada atributo por separado, debe ser alto.

Para llevar a cabo el análisis de conglomerados de *K* medias, se hicieron varios ensayos. Para cada uno se revisaron los tres criterios anteriores y a partir de tales condiciones se eliminaron algunas propuestas y se adoptó aquella que cumpliera estos criterios y pudiera

tener un significado a nivel educativo. Los conglomerados también sirven para establecer rutas de mejoramiento. En este documento se presentan las rutas de mejoramiento para los estudiantes que no dominan todos los atributos definidos para el pensamiento estadístico.

Para recopilar la información de las respuestas de los estudiantes en una base de datos se utilizó el programa Microsoft ® Excel (versión 2016). Para calcular la medida de acuerdo en la calificación (Kappa), para realizar los análisis descriptivos de la aplicación, así como los análisis de ítems se utilizó el software estadístico R (versión 3.4.2). Para los análisis de conglomerados y de regresiones múltiples se utilizó el software estadístico SPSS (versión 22). Para ejecutar los análisis del método *Rule Space* se utilizó el programa P-MAIN RULE SPACE (Tatsuoka, Varadi & Tatsuoka, 2004).

Resultados

Fase 1 – Desarrollo del Instrumento de Pensamiento Estadístico

Lista de componentes.

A partir de las discusiones desarrolladas en el panel de expertos sobre el dominio por evaluar, se adoptó una definición de Pensamiento Estadístico y se estableció que la faceta práctica sería la que se evaluaría con la prueba; esto se muestra en la Tabla 8.

Por otra parte, después de varias rondas de discusión sobre los componentes por evaluar, se llegó a una lista final con 14 elementos, que se muestra en la Tabla 8. Estos se definieron con base en los estándares del MEN y representan aquello que los estudiantes deben dominar al finalizar tercer grado. Los expertos elaboraron esta lista acompañada por la definición de cada uno de estos componentes.

C1 - Clasificar datos

Dado un conjunto de datos, agrupar elementos de acuerdo con características comunes dadas. Los datos deben poder organizarse en categorías que todos los niños dominen. Por ejemplo, clasificar por características cualitativas básicas como forma, tamaño, textura, entre otras. En el caso de utilizar criterios numéricos deben ser sencillos como pares, impares, decenas, conjuntos, etc.

Tabla 8

Definición y faceta del Pensamiento Estadístico.

Elemento	Definición
Dominio: Pensamiento Estadístico	El Pensamiento Estadístico es la capacidad que tienen los estudiantes para manejar datos presentados de manera gráfica y verbal con eficiencia, con el fin de extraer información útil y tomar decisiones con base en ella.
Faceta: Práctica	Expresa condiciones sociales de relación de las personas con su entorno, y contribuye a mejorar su calidad de vida y su desempeño como ciudadano.

Tabla 9

Lista final de componentes

Rótulo	Componente
C1	Clasificar datos
C2	Ordenar datos
C3	Describir características comunes de un conjunto de datos
C4	Identificar el patrón de un conjunto de datos
C5	Identificar tendencias de un conjunto de datos
C6	Representar un conjunto de datos mediante un diagrama de barras
C7	Representar un conjunto de datos mediante una tabla de frecuencias
C8	Representar un conjunto de datos mediante un pictograma
C9	Encontrar la correspondencia entre dos formas de representación de datos
C10	Interpretar lo que un diagrama de barras representa
C11	Interpretar lo que una tabla de frecuencias representa
C12	Interpretar lo que un pictograma representa
C13	Establecer conjeturas sobre la posibilidad de ocurrencia de ciertos eventos
C14	Predecir si la posibilidad de ocurrencia de un evento es mayor que la de otro

Evaluación de Pensamiento Estadístico con *Rule Space*

C2 - Ordenar datos

Dado un conjunto de datos, organizarlos de acuerdo con un criterio específico. Como criterios de ordenación pueden utilizarse variables cualitativas como tamaño u orden alfabético. Como criterio cuantitativo puede utilizarse la cantidad.

C3 - Describir características comunes de un conjunto de datos

Nombrar las características comunes de un conjunto de datos. Las características deben hacer referencia al grupo en su conjunto o a la mayor parte de él. Se debe indagar siempre por características comunes, no aquellas que diferencian los elementos del conjunto dado.

C4 - Identificar el patrón de un conjunto de datos

Encontrar la regla que sigue un conjunto de datos. El estudiante debe encontrar dicha regla en una secuencia de datos.

C5 - Identificar tendencias de un conjunto de datos

Predecir los valores que tomará un conjunto de datos a partir de su comportamiento. El estudiante puede predecir el siguiente elemento o grupo que pueden ser valores, diseños o formas.

C6 - Representar un conjunto de datos mediante un diagrama de barras

Construir un diagrama de barras a partir de información presentada, que puede ser textual, a partir de una tabla o de un pictograma. Debe usarse un máximo de tres categorías.

Evaluación de Pensamiento Estadístico con *Rule Space*

C7 - Representar un conjunto de datos mediante una tabla de frecuencias

Construir una tabla de frecuencias a partir de información presentada que puede ser textual, a partir de un diagrama de barras o de un pictograma. Tres categorías como máximo.

C8 - Representar un conjunto de datos mediante un pictograma

Construir un pictograma a partir de información presentada de manera textual, a partir de una tabla o de un diagrama de barras. Deben usarse máximo tres categorías.

C9 - Encontrar la correspondencia entre dos formas de representación de datos

Emparejar dos representaciones gráficas de un mismo conjunto de datos. El estudiante debe escoger una forma de representación análoga, es decir que presente la misma información.

C10 - Interpretar lo que un diagrama de barras representa

El estudiante debe comparar o extraer información básica sobre lo que se le presenta diagrama de barras.

C11 - Interpretar lo que una tabla de frecuencias representa

Comparar o extraer información básica sobre lo que se le presenta en la tabla de frecuencias.

C12 - Interpretar lo que un pictograma representa

Comparar o extraer información básica sobre lo que se le presenta en un pictograma.

Evaluación de Pensamiento Estadístico con *Rule Space*

C13 - Establecer conjeturas sobre la posibilidad de ocurrencia de ciertos eventos

Utilizar categorías como "improbable", "poco probable" y "muy probable" para describir la posibilidad de ocurrencia de ciertos eventos, con base en estimaciones.

C14 - Predecir si la posibilidad de ocurrencia de un evento es mayor que la de otro

Comparar la posibilidad de ocurrencia de dos eventos con base en estimaciones.

Especificaciones de la prueba de pensamiento estadístico.

Después da construir la lista de componentes del Pensamiento Estadístico para los estudiantes que finalizan grado tercero, uno de los resultados del panel de expertos fue la elaboración de las especificaciones de la prueba. Estas especificaciones se desarrollaron con base en los estándares para la construcción de pruebas de la APA, AERA y NCME (2014).

***Propósito de la prueba**:* Las puntuaciones de la prueba deben dar cuenta del desempeño específico de los estudiantes en el dominio del Pensamiento Estadístico. La finalidad principal de la prueba es establecer asignaciones de tipo (caracterizar), aunque también de nivel de desempeño. El campo de aplicación es el educativo.

***Especificaciones de contenido**:* La prueba busca dar cuenta del desempeño de los estudiantes y establecer un patrón de dominio (y de no dominio) de los componentes ya definidos. Estos componentes abarcan el campo del *Pensamiento Estadístico* en los grados primero a tercero.

Población*:* La población prevista para ser evaluada son los niños de primero a tercer grado de primaria. Sin embargo, teniendo en cuenta que los niños de primero apenas empiezan a recibir instrucción sobre los temas evaluados, es recomendable no utilizar la prueba con ellos. También pueden ser población objetivo los estudiantes de grado cuarto y quinto, para verificar si dominan los componentes evaluados.

Restricciones (tiempo, medios, situaciones): La prueba debe estar diseñada para ser aplicada en un tiempo relativamente corto (no sobrepasar una hora de aplicación). Debe estar diseñada para ser resuelta en lápiz y papel y el escenario definido para su aplicación será el aula de clase de los estudiantes, con esto se asegura la familiaridad de los estudiantes con el contexto de aplicación.

Especificaciones de formato: Se debe evitar el uso de hoja de respuestas ya que puede significar una carga adicional para los estudiantes. Debido a las características de los componentes por evaluar se utilizarán formatos de ítems de desarrollo (o de respuesta abierta) y de selección múltiple con única respuesta. Los ítems deberán ser presentados en un cuadernillo en el que los estudiantes marcarán la respuesta que consideren correcta o elaboren la respuesta (en un espacio determinado), según sea el caso. También, los expertos sugirieron que se evitara el uso excesivo de texto, debido a que la mayoría de los estudiantes están en proceso de aprendizaje de la lectura. Adicionalmente se recomendó utilizar imágenes ilustrativas para hacer llamativa la prueba para los participantes y facilitar la comprensión.

Especificaciones psicométricas**:** Dado el objetivo de la prueba y los componentes por evaluar, la dificultad de los ítems no es tan importante como la representatividad del dominio

evaluado, por lo que se hizo énfasis en los segundos. Además, teniendo en cuenta que se utilizará el modelo de Rasch en el análisis, es necesario que la prueba conserve la propiedad básica de independencia local. Al no existir elementos teóricos previos sobre la dimensionalidad del constructo, será algo que se comprobará empíricamente, mientras tanto, en la construcción de los ítems se tratará de asegurar que los ítems abarquen la evaluación de todos los componentes planteados.

Calificación de los ítems: La calificación de los ítems se realizará de manera dicotómica. En el caso de selección múltiple, habrá solamente una respuesta correcta (para la construcción de opciones de respuesta es que las incorrectas deben serlo en su totalidad, no se admitirán opciones parcialmente correctas). Para el caso de los ítems de respuesta construida o de desarrollo, se establecerán criterios de calificación estrictos donde se diferenciará claramente entre respuestas y tipos de respuestas correctas e incorrectas. Esto debe estar plasmado en la guía de calificación.

Aspectos culturales, de grupos minoritarios y estudiantes con discapacidad: Los ítems deben ser diseñados teniendo en cuenta la diversidad del contexto colombiano, tratando de asegurar que elementos de lenguaje, costumbres y relacionados puedan influir en el desempeño en la prueba. Teniendo en cuenta que la prueba incluye componentes de visualización de gráficos y dibujos, no es pertinente que sea presentada a estudiantes con discapacidad visual. Estudiantes con dificultades auditivas o del habla no tienen ninguna restricción. Por otra parte, los estudiantes con dificultades cognitivas podrán participar, a

menos que presenten algún trastorno que impida la comprensión de los enunciados. El colegio deberá entregar esta información previamente.

También, los expertos definieron una estructura de prueba que se presenta en la Tabla 10. Esta estructura ayudó a especificar las combinaciones de componentes admisibles en cada uno de los ítems, así como fijar la cantidad de ítems que conformarían la versión preliminar de la prueba.

Versión inicial de la prueba de pensamiento estadístico.

Como resultado del proceso de elaboración de ítems por parte del constructor de ellos y de la validación y corrección por parte del panel de expertos se obtuvo un conjunto de 24 ítems que conformaron la versión inicial de la prueba. Una tercera parte de la prueba consistió en ítems de selección múltiple, mientras que los demás ítems fueron de respuesta construida.

Para ofrecer criterios claros de calificación de los ítems de respuesta construida, los expertos elaboraron una guía de calificación que sirvió como guía para el equipo encargado de esta tarea. Esta guía fue mencionada anteriormente y se encuentra en el Anexo 6.

Tabla 10

Estructura de la Prueba de Pensamiento Estadístico.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
T	D	D	S	S	S	D	D	D	D	D	D	D	S	D	D	S	D	S	S	S	D	S	D	D
C1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
C2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
C3	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
C4	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
C6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
C7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
C8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
C9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
C10	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0
C11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0
C12	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C13	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
C14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Nota: La fila T hace referencia al tipo de formato propuesto para el ítem. En donde D son ítems de desarrollo o de respuesta construida y S son ítems de selección múltiple.

Aplicación piloto.

La aplicación piloto arrojó una adecuada aprehensibilidad de la prueba, es decir que los estudiantes entendieron la mayoría de los ítems. En aquellos en donde hubo dudas o algún tipo de ambigüedad se realizaron ajustes para agregar claridad a los textos, figuras o imágenes.

También se calculó el tiempo promedio de aplicación, que fue de 40 minutos. Sin embargo, hubo estudiantes que se demoraron alrededor de 30 minutos y otros que demoraron más de una hora en completar la prueba.

En resumen, la aplicación piloto sirvió para mejorar la presentación de los ítems y para preparar de manera adecuada las condiciones materiales de las aplicaciones subsiguientes. En la Tabla 11 se muestra específicamente las modificaciones que se hicieron a los ítems que presentaron algún tipo de problema en su construcción o presentación.

Las respuestas de los estudiantes fueron registradas en una base de datos para su posterior análisis psicométrico y con el método *Rule Space.* Las respuestas de los estudiantes en los ítems que fueron objeto de modificación no fueron tenidas en cuenta para los análisis posteriores.

También, las respuestas en los ítems de respuesta construida fueron utilizadas para enriquecer las categorías establecidas en la guía de calificación. La cual fue revisada con base en esta nueva información.

La versión final de la Prueba de Pensamiento Estadístico se muestra en el Anexo 8.

Tabla 11

Cambios aprobados por los expertos después de la aplicación piloto.

Ítem	Situación por mejorar	Cambio realizado
2	En el enunciado no era clara la instrucción acerca de qué se debía escribir la característica común elegida por el estudiante.	Se especificó en la instrucción que se debía hallar la característica común de las frutas y escribirla.
3	Para algunos estudiantes no era clara la idea de que las estrellas estaban organizadas de manera sucesiva.	Se incluyeron flechas entre las estrellas para hacer más clara la idea de que se presentaba una secuencia.
7	Los estudiantes completaban el pictograma, pero no se podía conocer cuál era el patrón que identificaban.	Se solicitó que incluyeran una explicación posterior al dibujo de los peces.
10	Varios estudiantes manifestaron no saber leer las horas en un reloj análogo.	Se reemplazaron las imágenes de un reloj análogo por imágenes de un reloj digital.
11b	Cuando se pedía completar un pictograma, con dibujos de ranas, a muchos estudiantes se les dificultaba dibujarlas.	Se cambió la instrucción para que los estudiantes dibujaran gotas de agua.
13	El ítem solicita que se identifique una característica común entre un grupo de figuras geométricas. Los estudiantes las dibujaban, pero en ocasiones no se podía conocer el criterio de agrupación.	Se pidió que enunciaran la característica escogida como criterio para formar el conjunto.
Prueba	Según algunos docentes y estudiantes, el tipo de letra podría cambiarse y ser un poco más atractivo. Los márgenes eran muy amplios y la letra era pequeña.	Se hicieron los márgenes más estrechos, de manera que el tamaño de la letra se aumentó a 14 y el tipo se cambió a "Comic Sans" (letra más atractiva).

Fase 2 – Implementación del Rule Space

Aplicación del Instrumento.

La prueba de pensamiento estadístico fue aplicada a un total de 1580 estudiantes de las ciudades de Bogotá, La Paz, Valledupar, Tumaco y San José del Guaviare entre los meses de mayo de 2017 y septiembre de 2018. Sin embargo, la muestra fue reducida a 1235 estudiantes debido a que aquellos casos que presentaron un porcentaje mayor al 30% de omisiones fueron descartados.

De los 1235 estudiantes, 595 eran niñas, 638 niños y 2 personas no entregaron información sobre su sexo; con un rango de edades entre 7 y 15 años (M=9.11, DE=1.096). La cantidad de estudiantes por grado fue: 161 de segundo (80 niñas, 81 niños), 381 de tercero (181 niñas, 200 niños), 582 de cuarto (286 niñas, 296 niños) y 108 de quinto (48 niñas y 60 niños). En la Tabla 12 se muestra la cantidad de participantes por ciudad, colegio y tipo de institución.

El resultado de la medida de consenso entre las dos personas responsables de la calificación de las pruebas fue satisfactorio (Kappa = 0.896, alfa = 0.000). Lo que indica una aplicación adecuada de los criterios de puntuación, así como claridad en la definición de estos.

Tabla 12

Cantidad de estudiantes por ciudad, colegio y tipo de colegio en la aplicación de la prueba.

Ciudad	**Colegio**	**Tipo**	
		No oficial	**Oficial**
Bogotá	Altamira Sur Oriental		345
	Calasanz	190	
	Colegio Sagrado Corazón	192	
	Instituto Infantil Y Juvenil	167	
La Paz	Ciro Pupo Martinez		95
SJG	CDR		184
Tumaco	ITPC		15
Valledupar	CDV		47
Total		**549**	**686**

Análisis psicométricos.

La puntuación media fue de 13.87 (DE=4.15), con un mínimo de 3 y un máximo de 24, lo que significa que, en promedio, los estudiantes tuvieron un desempeño cercano al 60% en la prueba.

Para explorar las diferencias en las puntuaciones en la prueba entre los diferentes factores sociodemográficos se corrieron pruebas de diferencias de medias no paramétricas (se ejecutó la prueba de Kolmogorov-Smirnov, encontrando que la distribución de las puntuaciones difería de una distribución normal (KS = 0.065, p = 0.000)). Los resultados de este ejercicio se muestran en la Tabla 13.

Tabla 13

Diferencias de medias, tomando diferentes factores como referencia.

Factor	Niveles	N	Media	Test	Valor	Sig.	Grupos diferentes
Género	Femenino	595	13.76	M-W	183590	0.319	
	Masculino	638	13.99				
Grado	Segundo	161	12.20	K-W	71.332	0.000	2-3* 2-4**, 2-5**, 3-4**, 3-5**
	Tercero	382	13.11				
	Cuarto	583	14.72				
	Quinto	108	14.50				
Ciudad	Bogotá	894	14.17	K-W	49.819	0.000	LPZ-TUM*, BOG-VUP*, SJG-VUP* BOG-LPZ**, LPZ-SJG**, BOG-TUM**, SJG-TUM**, VUP-TUM**
	La Paz	95	11.89				
	SJG	184	14.22				
	Tumaco	15	8.80				
	Valledupar	47	12.53				
Tipo	Oficial	686	12.68	M-W	118870.5	0.000	
	No oficial	549	15.36				
Colegio	IIJ (BOG)	167	14.61	K-W	189.42	0.000	6-7*, 5-8* 1-2**, 1-3**, 2-3**, 2-4**, 3-4**, 1-5**, 2-5**, 3-5**, 4-5**, 2-6**, 3-6**, 5-6**, 1-7**, 3-7**, 4-7**, 5-7**, 1-8**, 3-8**, 4-8**, 6-8**
	ASO (BOG)	345	12.27				
	Calasanz (BOG)	190	16.50				
	CSC (BOG)	192	14.88				
	ITPC (TUM)	15	8.80				
	CDR (SJG)	184	14.22				
	CDV (VUP)	47	12.53				
	CPM (LPZ)	95	11.89				

M-W: Mann-Withney K-S: Kruskal-Wallis * $p < 0.05$ ** $p < 0.01$

De acuerdo con lo que se observa en la Tabla 13 no se presentaron diferencias entre niños y niñas. En cuanto al factor de grado, se da lo esperado en la medida en que entre cuarto y quinto no hay diferencias significativas (se supone que ya dominan los componentes definidos para tercero), mientras que sí se presentan diferencias significativas en la media en la prueba entre estos dos grados y entre tercero y segundo (que están en proceso de dominar los atributos). No se presentan diferencias significativas entre Bogotá y San José del Guaviare y entre La Paz y Valledupar (estos últimos se encuentran en la misma zona metropolitana); Tumaco tuvo el peor desempeño, aunque hay que considerar que solo se pudieron conservar 15 estudiantes debido a la gran cantidad de omisiones. Por otra parte, los colegios no oficiales tuvieron un mejor desempeño que los oficiales; considerando que entre los primeros se encuentra el Calasanz, que tuvo el mejor desempeño entre los participantes y en el segundo grupo está el ITPC de Tumaco que tuvo el peor desempeño.

También se exploraron las características técnicas de los ítems y se revisaron especialmente los criterios definidos con anterioridad para eliminar ítems que presentaran problemas. En la Tabla 14 se muestran las características de los ítems respecto a la correlación ítem-prueba (C1), medida de consistencia interna si se elimina el ítem (C2), infit (C3) y outfit (C4).

Tabla 14

Características técnicas de los ítems

Ítem	C1	C2	C3	C4	Ítem	C1	C2	C3	C4
1	0.14	0.68	1.00	1.02	13	0.36	0.68	0.76	0.58
2	0.31	0.66	0.98	0.96	14	0.23	0.68	1.06	1.07
3	0.03	0.68	1.20	1.39	15	0.19	0.68	0.98	1.11
4	0.22	0.67	1.04	1.01	16	0.19	0.68	1.08	1.13
5	0.17	0.68	1.11	1.12	17	0.39	0.67	0.98	1.03
6	0.43	0.66	0.86	0.80	18	0.37	0.66	0.87	0.76
7	0.15	0.68	1.00	1.24	19	0.20	0.67	1.07	1.06
8	0.23	0.68	0.86	0.86	20	0.37	0.67	0.88	0.84
9	0.40	0.66	0.90	0.88	21	0.52	0.65	0.80	0.75
10	0.19	0.67	1.05	1.14	22	0.33	0.67	0.96	0.86
11	0.34	0.67	0.83	0.70	23	0.51	0.65	0.80	0.76
12	0.31	0.67	0.84	0.69	24	0.49	0.65	0.83	0.78

Después de realizar la verificación de esta información, se decidió eliminar los ítems 3, 15, 16 y 17: El ítem 3 mostró una correlación casi nula con la puntuación total en la prueba (0.03), además de desajuste lejano y cercano. El ítem 13 tuvo una correlación muy baja con el total de la prueba (0.06), además de desajuste lejano y cercano. El ítem 14 tuvo correlación muy baja con el total de la prueba (0.19), además de desajuste lejano y cercano. Por último, el ítem 15 mostró una **c**orrelación muy baja con el total de la prueba (0.19), así como desajuste lejano y cercano. Los resultados que siguen se realizaron con la versión de prueba de 20 ítems.

Evaluación de Pensamiento Estadístico con *Rule Space*

El desempeño de los estudiantes en la prueba depurada se presenta como la suma o el total de ítems acertados. La puntuación mínima posible es 0 y la máxima es 20. La distribución de la proporción de aciertos se muestra en la Figura 6. Se observa que el 25% de los estudiantes contestó correctamente 9 ítems o menos, el 50% acertó 12 ítems o menos y, finalmente, el 75% contestó correctamente 15 ítems o menos.

Para evaluar la fiabilidad del nuevo conjunto de 20 ítems se utilizó el alfa de Cronbach, encontrando una consistencia interna adecuada, a = .75, IC 95% [.73, .75]. Además, se observó que ningún ítem afectaba considerablemente la confiabilidad en caso de ser eliminado, por lo que se toma la decisión de no eliminar más.

El índice de dificultad en Teoría Clásica para los ítems varió de 0.169 a 0.925 con una media aritmética igual a 0.599 y una desviación estándar de 0.209. En cuanto a la discriminación, medida con la correlación biserial puntual, los valores oscilaron entre 0.380 y 0.777 con un promedio de 0.565 y una desviación estándar de 0.134.

Según los resultados arrojados al utilizar el modelo de Rasch, el parámetro de dificultad de los ítems varió de -2.13 a 2.60. El ítem con mayor dificultad fue el 7 y el más fácil fue el ítem 8. Se observan 11 ítems por encima y 9 por debajo de la dificultad media. Esto se puede observar en la Figura 7 donde se presenta el mapa de ítems y personas. El error estándar para el cálculo del parámetro de dificultad osciló entre 0.09 y 1.17.

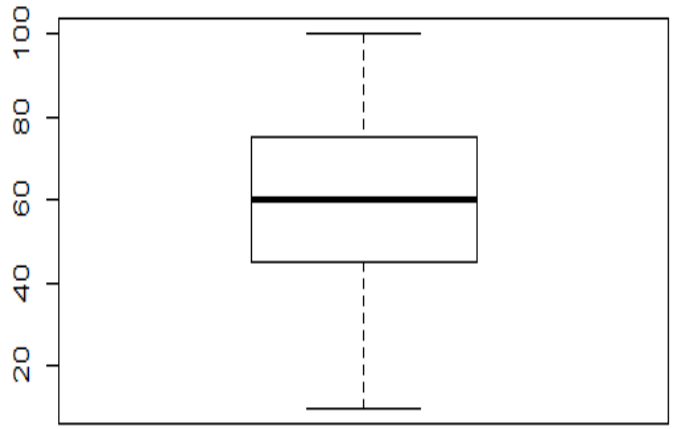

Figura 6. Distribución de la proporción de aciertos de la muestra.

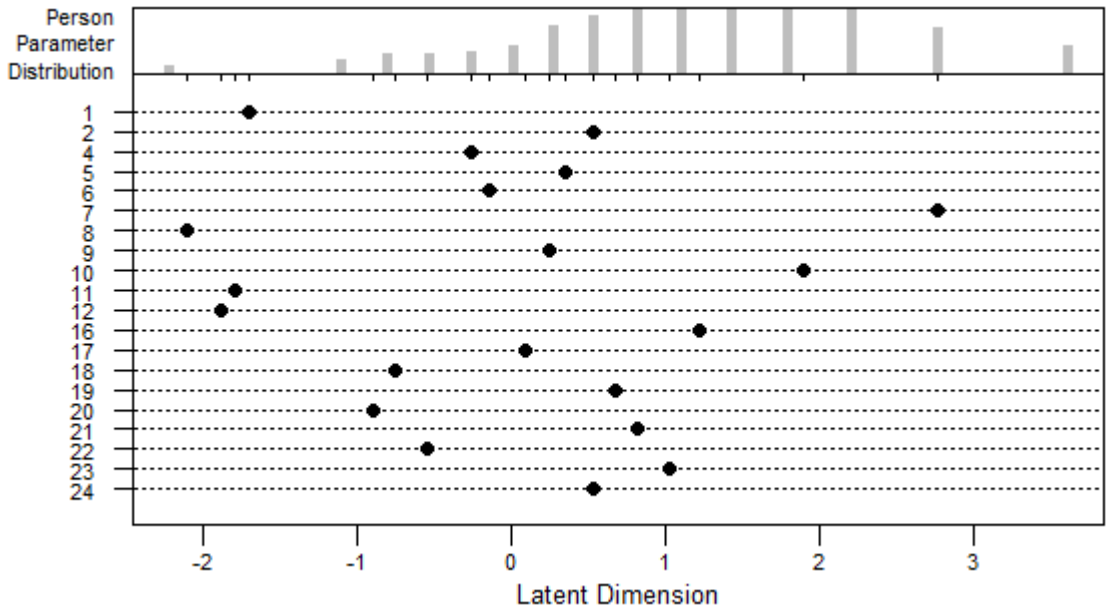

Figura 7. Mapa de personas e ítems. Calculado con el modelo de Rasch.

En cuanto a las medidas de ajuste interno (INFIT) y externo (OUFIT) se observa que ningún ítem estuvo significativamente por fuera del rango de ajuste aceptable sugerido por Wright y Linacre (1994). Entre los ítems que estuvieron más cerca de desajustar está el 11 con un outfit MSQ de 0.670 y un infit MSQ de 0.859 y el ítem 17 que muestra un outfit MSQ de 1.299 y un infit MSQ de 1.066. En esta etapa no fue necesario suprimir ningún ítem del análisis debido a que ninguno presentó valores de desajuste extremos.

Rule Space – Identificación de atributos.

Los atributos definidos a partir de la lista de componentes se muestran en la Tabla 15. Al haber menos atributos que componentes definidos para la construcción de la prueba, los atributos se hacen un poco menos granulares.

Tabla 15

Lista de atributos.

Rótulo	Atributo
AT1	Ordenar datos
AT2	Identificar características comunes dentro de un conjunto de datos
AT3	Completar secuencias
AT4	Identificar el patrón dentro de un conjunto de datos
AT5	Interpretar situaciones descritas en el enunciado
AT6	Estimar la probabilidad de ocurrencia de ciertos eventos
AT7	Determinar frecuencias dentro de un conjunto de datos
AT8	Representar gráficamente información dada
AT9	Identificar el elemento mínimo y/o máximo de un conjunto
AT10	Extraer información a partir de representaciones gráficas

Evaluación de Pensamiento Estadístico con *Rule Space*

Para verificar si la cantidad y la configuración de atributos en cada ítem pueden explicar la varianza en la dificultad, se ejecutaron análisis de regresión múltiple. Para ello, se utilizó el valor de dificultad (parámetro beta), calculado con el modelo de Rasch, como variable dependiente y los vectores binarios de atributos como variables independientes. En un siguiente análisis de regresión múltiple se tomó como variable dependiente la dificultad de los ítems, pero calculada con el modelo de la Teoría Clásica de los Tests. Los resultados se muestran en la Tabla 16.

Tabla 16

Regresión múltiple respecto a la dificultad explicada por la configuración de atributos en cada ítem.

Modelo	**R^2**	**R^2 ajustado**
Rasch	0.868	0.721
Teoría Clásica de los Tests	0.869	0.723

Lo anterior indica que la lista y la configuración de atributos propuesta es capaz de explicar la dificultad de los ítems y por lo tanto el desempeño de los evaluados en la prueba.

Rule Space - Construcción de matrices.

Debido a que en este caso se trató de una especie de ajuste posterior, en la Figura 8 se muestran las matrices A y R tal como quedarían definidas dado que no se ha asumido ninguna relación de dependencia entre los atributos especificados.

$$A=\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad R=\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Figura 8. Matrices A y R.

La matriz A indica que no existe una relación de dependencia directa entre los atributos. Mientras que la matriz R indica que tampoco hay relaciones de dependencia indirecta, pero sí de cada atributo con sí mismo.

Teniendo en cuenta que existen 10 atributos, sin ninguna relación de dependencia ni teórica ni demostrada empíricamente, la matriz Q tendría $2^{10} - 1$ combinaciones de atributos, es decir 1023 columnas. Sin embargo, debido a que para esta situación concreta se realizó una recodificación de los ítems con base en los nuevos atributos, la matriz Q (que aquí se nombró como matriz Q ajustada) corresponde al resultado de la codificación de los ítems de la Prueba de Pensamiento Estadístico. Esta matriz se muestra en la Figura 9.

$$Qa = \begin{array}{c|cccccccccccccccccccccc} & 1 & 2 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\ AT1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ AT2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ AT3 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ AT4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ AT5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ AT6 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ AT7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ AT8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ AT9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ AT10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}$$

Figura 9. Matriz Q ajustada.

Para verificar la representatividad de los atributos dentro de la prueba se realizó un conteo de cada uno de ellos dentro del cuerpo de ítems. En la Tabla 17 se muestra la cantidad de veces que cada atributo es evaluado en la prueba de pensamiento estadístico.

Tabla 17

Cantidad de ítems que requieren de cada atributo.

Atributo	Frecuencia
AT1	3
AT2	3
AT3	3
AT4	2
AT5	6
AT6	5
AT7	5
AT8	4
AT9	1
AT10	4

Evaluación de Pensamiento Estadístico con *Rule Space*

Los atributos que más se evalúan en el cuerpo de ítems de la Prueba de Pensamiento Estadístico son el AT5 (Interpretar situaciones descritas en el enunciado) que aparece en 6 ítems y los atributos AT6 (Estimar la probabilidad de ocurrencia de ciertos eventos) y AT7 (Determinar frecuencias dentro de un conjunto de datos) que aparecen en 5 ítems cada uno.

En la tabla 18 se muestra el número de atributos que corresponden a cada ítem de la prueba, y el parámetro de dificultad asociado. Allí se observa que los ítems evalúan entre uno y cuatro atributos. En la literatura sobre *Rule Space* se asume que los ítems que impliquen un mayor número de atributos para su resolución serán los más difíciles, pero parece que este no es el caso. Por ejemplo, el ítem 17 requiere de cuatro atributos para su resolución, pero su nivel de dificultad es de apenas 0.09. El ítem 11 requiere de tres atributos para su resolución y tiene una dificultad de -1.80. Por otra parte, los ítems 2 y 24 solo requieren de un atributo para ser resueltos y tienen niveles de dificultad de 0.53 y 0.52, respectivamente. Es posible que, en este caso, la dificultad de los ítems este siendo explicada por *cuáles* atributos están implicados en ellos y no por *cuántos*. Por ejemplo, se observa que ítems con parámetros de dificultad mayor a 1.00 como el 7, el 10, el 16 y el 23 requieren los atributos AT4 (Identificar el patrón dentro de un conjunto de datos) o AT5 (Interpretar situaciones descritas en el enunciado).

Tabla 18

Cantidad de atributos que requiere cada ítem

Ítem	Frec.	Dificultad (Rasch)	Atributos	Ítem	Frec.	Dificultad (Rasch)	Atributos
1	1	-1.71	1000000000	12	2	-1.89	0000000011
2	1	0.53	0100000000	16	2	1.22	0000110000
4	1	-0.26	0000010000	17	4	0.09	1100001100
5	2	0.35	0000110000	18	1	-0.76	0000000001
6	1	-0.15	0010000000	19	2	0.67	0010100000
7	2	2.76	0011000000	20	1	-0.90	0000000001
8	1	-2.11	0000001000	21	3	0.81	0000101100
9	3	0.25	0100001100	22	1	-0.55	0000010000
10	2	1.89	0001100000	23	2	1.03	1000100000
11	3	-1.80	0000001101	24	1	0.52	0000010000

Rule Space - Clasificación de los evaluados.

Entre los evaluados que conformaron la muestra para este análisis, lograron ser clasificados 557 en algún estado de conocimiento, lo que representa una tasa de clasificación del 99.1%. Esta tasa de clasificación es considerada muy alta (Chen, 2006; Dogan, 2006; Artavia-Medrano, 2014).

Las cinco personas que no pudieron ser clasificadas según el R*ule Space*, se identificaron con los números 67, 174, 417, 464 y 533. Los cuatro primeros estuvieron por encima del criterio establecido como distancia máxima con el centroide más cercano y el último no pudo

ser clasificado en ningún estado de conocimiento por el modelo. En la Tabla 19 se muestra el patrón de respuestas de las cinco personas no clasificadas según el *Rule Space* y la cantidad total de ítems acertados en la prueba.

La razón por la cual en *Rule Space* estas personas no pudieron ser clasificadas se debe a que mostraron patrones de respuesta inconsistentes con los estados de conocimiento generados por el software, por lo que se ubican muy lejos a cualquier patrón ideal de respuesta esperados de acuerdo con el modelo. Estas discrepancias pueden deberse a errores sistemáticos o aleatorios por parte de los evaluados al momento de resolver la prueba.

El método *Rule Space* ofrece un cálculo de las probabilidades de dominio de atributos para cada estudiante clasificado en algún estado de conocimiento. Por ello, los resultados se presentan en adelante tuvieron en cuenta como muestra a las 557 personas clasificadas en algún estado de conocimiento.

En diferentes estudios con *Rule Space* se ha utilizado la probabilidad de dominio de atributos como una forma de validación de la matriz Q (Dogan, 2006; Chen, 2006; Artavia-Medrano, 2014). Por este motivo se realizó un análisis de regresión múltiple con la puntuación total en la prueba como variable dependiente y las probabilidades de dominio de atributos como variables independientes. Se obtuvo un R^2 igual a 0.976 y un R^2 ajustado igual a 0.975; esto indica que el 98% de la varianza en las puntuaciones totales de la prueba se explican por las probabilidades de dominio de atributos. Esto brinda otra evidencia de que la matriz Q propuesta es válida.

Tabla 19

Personas no clasificadas en el Rule Space.

ID	Patrón observado de respuestas	Total de aciertos
67	01111111111101111111	18
174	11111111111010000111	15
417	10001011001011101101	11
464	10110100011010000100	8
533	11011110111110011100	14

En la tabla 20 se muestran los promedios y las desviaciones estándar de la probabilidad de dominio de los 10 atributos evaluados (organizados de menor a mayor según probabilidad de dominio, es decir, del más "difícil" al más "fácil).

Como se observa en la Tabla 20, el atributo con la probabilidad promedio más baja de dominio es el AT4 (Identificar el patrón dentro de un conjunto de datos) con 0.367 los demás se encuentran por encima de 0.6 y dos en particular, AT7 (Determinar frecuencias dentro de un conjunto de datos) y AT10 (Extraer información a partir de representaciones gráficas), tienen probabilidad promedio muy cercana a 1. Estos últimos pueden verse como aquellos que son una fortaleza para los estudiantes evaluados. Esto se aprecia en más detalle en la Figura 10.

Tabla 20

Promedio y desviación estándar de las probabilidades de dominio de atributos para la muestra.

Atributo	Promedio	DE
AT4	0.367	0.312
AT3	0.666	0.278
AT9	0.695	0.231
AT5	0.727	0.374
AT2	0.737	0.282
AT1	0.829	0.245
AT6	0.841	0.287
AT8	0.867	0.248
AT10	0.980	0.110
AT7	0.980	0.111

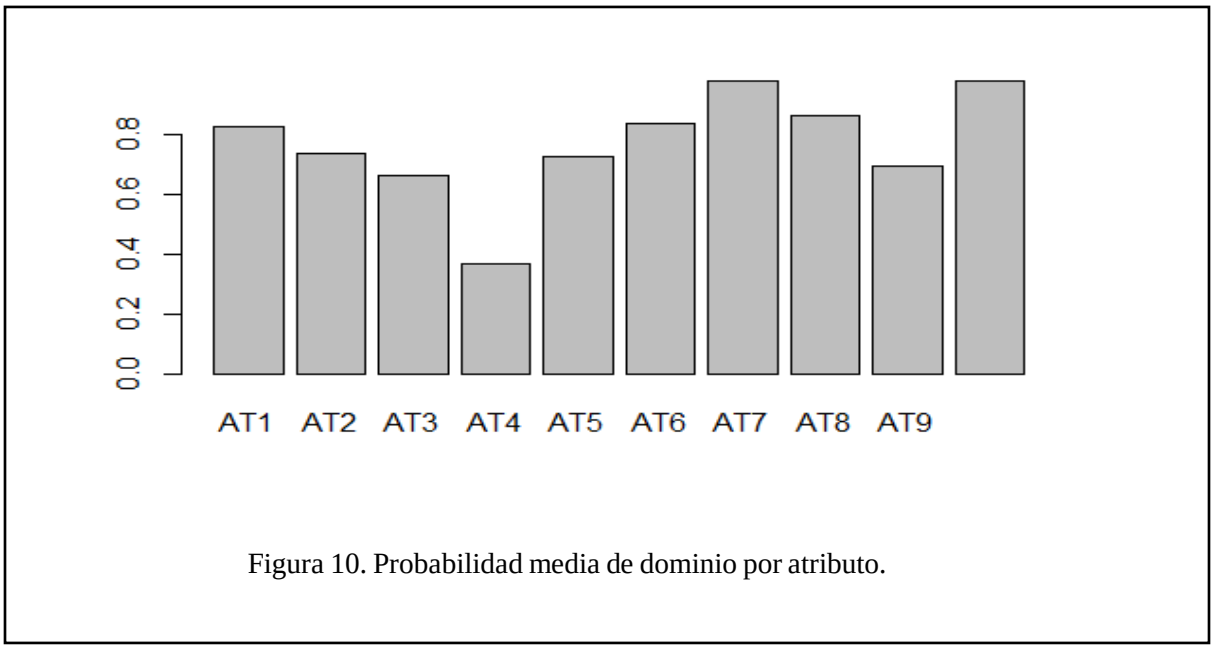

Figura 10. Probabilidad media de dominio por atributo.

Evaluación de Pensamiento Estadístico con *Rule Space*

El método *Rule Space*, por medio de la función booleana descriptiva, permite generar todas las posibles combinaciones de patrones de atributos, que pueden ser diferentes incluso en estudiantes con el mismo puntaje en la prueba. En la Tabla 21 se muestra un ejemplo de cómo tres personas con el mismo puntaje en la prueba muestran un patrón ideal de atributos diferente.

Tabla 21

Ejemplo de patrones ideales de atributos diferentes, en estudiantes con el mismo puntaje en la prueba.

ID	Puntaje asociado x/100	Habilidad (Rasch)	Patrón ideal de atributos asociado
114	35	0.816	1110011000
412	35	0.816	0001101111
463	35	0.816	0011001111

La información presentada en la Tabla 21 muestra diferencias entre los patrones ideales de atributos de las tres personas identificadas como 114, 412 y 463. Esto se presenta a pesar de contar con un mismo puntaje en la prueba y un mismo nivel de estimación de habilidad según el modelo de Rasch. En el caso de 114 se observa que le falta dominar los atributos AT4 (Identificar el patrón dentro de un conjunto de datos), AT5 (Interpretar situaciones descritas en el enunciado), AT8 (Representar gráficamente información dada), AT9 (Identificar el elemento mínimo y/o máximo de un conjunto) y AT10 (Extraer información a partir de representaciones gráficas); el patrón del estudiante con id 412 indica que le falta dominar AT1 (Ordenar datos), AT2 (Identificar características comunes dentro de un conjunto de datos),

AT3 (Completar secuencias) y AT6 (Estimar la probabilidad de ocurrencia de ciertos eventos); por último, el estudiante 463 no domina AT1 (Ordenar datos), AT2 (Identificar características comunes dentro de un conjunto de datos), AT5 (Interpretar situaciones descritas en el enunciado) y AT6 (Estimar la probabilidad de ocurrencia de ciertos eventos).

Los patrones ideales de atributos se generan a partir de los patrones observados de respuesta y las especificaciones de atributos dadas en la matriz de incidencia. Esta información es clave, ya que ofrece el detalle acerca de los atributos que domina cada estudiante y aquellos que está en proceso de dominar. Es aquí donde aparece el valor diagnóstico y su repercusión en el plano educativo, debido a que se entrega información pormenorizada para mejorar el proceso de enseñanza-aprendizaje.

En la Tabla 22 se muestra el estado de conocimiento más próximo en que cada persona de la Tabla 18 se ha clasificado, así como el valor del índice de precaución o atipicidad (segunda coordenada del R*ule Space*).

Tabla 22

Estado de conocimiento más próximo y su ubicación en el Rule Space.

ID	**EC más próximo**	**Valor de ζ**	**Patrón ideal de respuestas asociado** **Patrón observado de repuestas**
114	51	1.19	1 1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0
412	461	0.11	0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 0
463	433	0.12	0 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 1 0 0

Una vez más se resalta cómo personas que tienen la misma puntuación entre sí o la misma habilidad según el modelo de Rasch, pueden tener distintas caracterizaciones de su desempeño en función de los atributos cognitivos que subyacen a la prueba. Asimismo, el valor de ζ reconoce como muy inusual el patrón de respuesta de la persona identificada como 114. En la Tabla 23 se presentan los estados más comunes en los que se clasificaron las personas de la muestra de esta investigación.

Para construir la Tabla 23 se consideraron aquellos estados de conocimiento que clasificaron al menos al 1% de la muestra y para cada uno de ellos se muestra cuáles son los atributos que están en proceso de dominar por parte de las personas clasificadas en ellos.

En la Figura 11 se muestra la representación gráfica de los EC representados en la tabla 22, en donde el eje x representa el parámetro de atipicidad y el eje y representa el parámetro de dificultad del modelo de Rasch. Allí se observa una mayor concentración de estados de conocimientos con parámetro de habilidad por encima de 1, lo que es consistente con lo descrito anteriormente respecto a la habilidad media de las personas descrita anteriormente.

Tabla 23

Estados de conocimiento más frecuentes ordenados según el valor de θ.

EC	θ	ζ	Frecuencia	%	Atributos no dominados
1	4.000	0.600	12	2.154	-
2	3.608	1.117	23	4.129	AT9
10	2.759	-1.485	40	7.181	AT4
11	2.208	0.619	34	6.104	AT4, AT9
19	2.208	-0.005	17	3.052	AT3
127	2.208	1.789	9	1.616	AT2
20	1.781	0.922	8	1.436	AT3, AT9
22	1.781	-0.661	49	8.797	AT3, AT4
23	1.421	0.261	13	2.334	AT3, AT4, AT9
64	1.421	2.221	6	1.077	AT6
136	1.421	-0.357	27	4.847	AT2, AT4
256	1.421	-0.113	13	2.334	AT1, AT4
13	1.102	0.386	9	1.616	AT4, AT8
37	1.102	-0.373	36	6.463	AT5
137	1.102	0.587	5	0.898	AT1, AT3, AT9
145	1.102	0.926	7	1.257	AT2, AT3
38	0.809	0.421	5	0.898	AT5, AT9
46	0.809	-1.807	13	2.334	AT4, AT5
82	0.809	-0.073	17	3.052	AT4, AT6
151	0.809	-0.040	11	1.975	AT2, AT3, AT4
268	0.809	0.412	5	0.898	AT1, AT3.AT4
28	0.533	0.705	5	0.898	AT3.AT4, AT8

47	0.533	-0.832	14	2.513	AT4, AT5, AT9
55	0.533	-1.507	11	1.975	AT3, AT4, AT5
84	0.533	0.932	8	1.436	AT4, AT6, AT9
283	0.533	0.489	5	0.898	AT1, AT5
163	0.265	-0.743	18	3.232	AT2, AT5
104	0.264	-1.108	6	1.077	AT3, AT4, AT5, AT6
83	0.001	-1.201	5	0.898	AT4, AT5, AT6

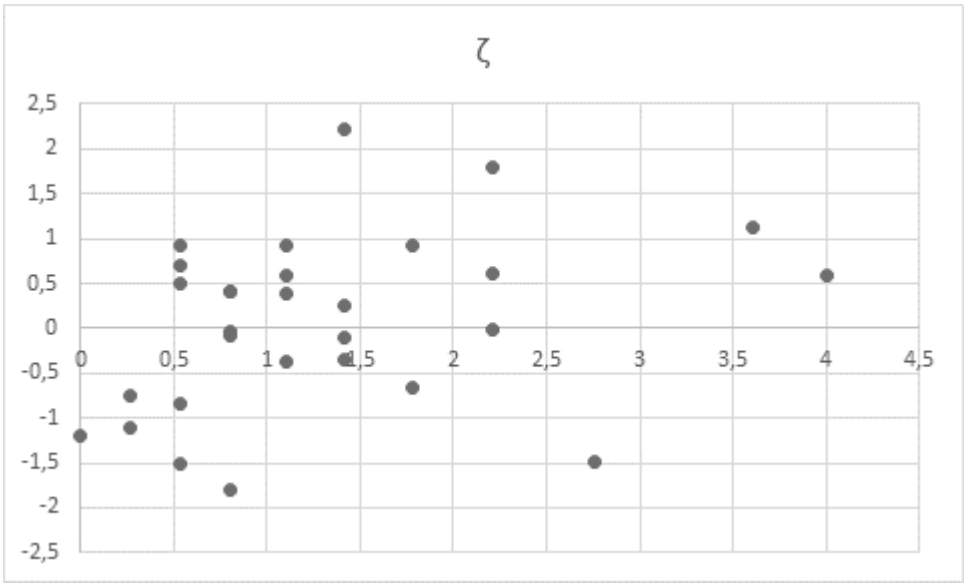

Figura 11. Representación en el espacio bidimensional de los estados de conocimiento más frecuentes.

Rule Space - Estados conglomerados de conocimiento.

El modelo encontró 480 estados ideales, lo que representa un 46.9% de todos los estados de conocimiento posibles. Sin embargo, la cantidad de EC fue muy alta para poder realizar un análisis a nivel colectivo que se útil. Para ello, se realizó una agrupación de los estados de conocimiento utilizando un análisis de conglomerado de K medias.

En este análisis pueden seleccionarse la cantidad de agrupaciones a discreción, pero en este caso se siguieron las recomendaciones realizadas por Chen, Gorin, Thompson y Tatsuoka (2008). Además de cumplir los criterios planteados por esos autores, se seleccionó la cantidad de 9 conglomerados por considerar que así se refleja mejor el desempeño o logro de los atributos y porque de estos conglomerados es posible obtener una relación jerárquica cuya interpretación tenga mayor sentido.

Los 9 conglomerados se nombraron como ECC1 a ECC9. La cantidad de atributos dominados en cada estado conglomerado de conocimiento se presenta en la tabla 24.

Con el propósito de entender la importancia de cada atributo en la separación por los nueve conglomerados, se llevó a cabo un análisis de varianza (ANOVA) de un solo factor. Los resultados que se presentan en la Tabla 24, indican cuáles atributos contribuyen más a la solución escogida de los estados conglomerados de conocimiento. Así, los atributos con valores de *F* grandes proporcionan mayor separación entre conglomerados; en este caso, AT5 (Interpretar situaciones descritas en el enunciado).

Tabla 24

Cantidad de atributos dominados en cada estado conglomerado de conocimiento.

Estados conglomerados de conocimiento	Cantidad de atributos dominados
ECC4	6
ECC1	5
ECC3	4
ECC6	4
ECC2	3
ECC7	3
ECC5	2
ECC8	1
ECC9	1

Puesto que los conglomerados se han elegido para maximizar las diferencias entre los casos, las pruebas *F* solo se pueden utilizar con una finalidad descriptiva, esto es, no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

Los resultados presentados en la Tabla 25 muestran que los valores de *F* varían desde 21.316 hasta 361.369, esto denota que cada atributo es de utilidad para la separación en estados conglomerados de conocimiento. Por lo tanto, la solución de nueve conglomerados se considera una solución interpretable en términos cognitivos para los propósitos de este estudio. En este caso se evidencia que la habilidad para poder interpretar el texto en los ítems puede ser determinante en el desempeño en la prueba de pensamiento estadístico.

Tabla 25

Resultados del ANOVA de un solo factor para evaluar las diferencias entre los doce estados conglomerados de conocimiento por atributo.

Atributo	**Conglomerado**	**Error**	**F**
	Media cuadrática (gl = 8)	**Media cuadrática (gl = 548)**	
AT1	1.673	0.037	45.546
AT2	2.216	0.048	45.803
AT3	2.301	0.045	51.156
AT4	3.823	0.043	89.122
AT5	8.186	0.023	361.639
AT6	1.358	0.064	21.316
AT7	0.480	0.006	86.795
AT8	2.477	0.026	94.082
AT9	1.115	0.038	29.530
AT10	0.486	0.005	91.494

Una vez que se obtuvieron los conglomerados, se estimó la probabilidad de dominio o logro para cada atributo en cada conglomerado, mediante el promedio de las probabilidades de dominio de atributos para las personas clasificadas en el conglomerado respectivo. En la Tabla 26 se muestra la probabilidad media y la cantidad de personas en cada conglomerado.

Tabla 26

Centros de cada estado conglomerado de conocimiento.

	Estados Conglomerados de Conocimiento (ECC)								
Atributo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
AT1	0.96	0.63	0.72	0.98	0.66	0.49	0.87	0.54	0.67
AT2	0.83	0.60	0.59	0.89	0.32	0.58	0.87	0.63	0.50
AT3	0.24	0.37	0.69	0.80	0.50	0.62	0.76	0.67	0.85
AT4	0.39	0.66	0.05	0.57	0.20	0.85	0.27	0.21	0.20
AT5	0.97	0.97	0.95	0.94	0.27	0.91	0.14	0.42	0.29
AT6	0.90	0.97	0.92	0.91	0.45	0.86	0.83	0.29	0.65
AT7	1.00	0.53	1.00	1.00	1.00	0.99	1.00	0.92	0.64
AT8	0.96	0.03	0.87	0.96	0.72	0.92	0.91	0.29	0.16
AT9	0.55	0.40	0.76	0.66	0.87	0.88	0.74	0.12	0.50
AT10	0.97	0.93	0.99	1.00	0.99	1.00	0.99	0.29	0.98
Promedio	0.78	0.61	0.75	0.87	0.60	0.81	0.74	0.44	0.54
Cantidad de casos	64	10	116	168	41	36	100	8	14

Para facilitar la interpretación se realizó una transformación de las probabilidades medias de dominio de atributos para cada uno de los ECC, estableciendo un patrón binario, de manera que se pudiera definir si se domina o no cada uno de los atributos. Se utilizó 0.75 como punto de corte para estas transformaciones. Esto se muestra en la Tabla 27.

Tabla 27

Transformación binaria de atributos dominados para cada uno de los ECC definidos.

	Estados Conglomerados de Conocimiento (ECC)								
Atributo	1	2	3	4	5	6	7	8	9
AT1	1	0	0	1	0	0	0	0	0
AT2	0	0	0	0	0	0	0	0	0
AT3	0	0	0	0	0	0	0	0	0
AT4	0	0	0	0	0	0	0	0	0
AT5	1	1	1	1	0	1	0	0	0
AT6	0	1	1	1	0	0	0	0	0
AT7	1	0	1	1	1	1	1	1	0
AT8	1	0	0	1	0	1	1	0	0
AT9	0	0	0	0	0	0	0	0	0
AT10	1	1	1	1	1	1	1	0	1
Atributos dominados	5	3	4	6	2	4	3	1	1
N clasificados	64	10	116	168	41	36	100	8	14

Se evidencia que la mayoría de las personas clasificadas (168) se ubican en el estado conglomerado de conocimientos 4 (ECC4). Las personas que se ubican en este conglomerado no dominan los atributos AT2 (Identificar características comunes dentro de un conjunto de datos), AT3 (Completar secuencias.), AT4 (Identificar el patrón dentro de un conjunto de datos) y AT9 (Identificar el elemento mínimo y/o máximo de un conjunto). En el siguiente ECC con mayor cantidad de clasificados, el ECC 3, se presentan los mismos atributos no

dominados más el AT1 (Ordenar datos) y el AT8 (Representar gráficamente información dada).

En la Figura 12 se presenta la jerarquía propuesta entre los ECC. Según esta jerarquía, los ECC más básicos son el ECC8 y el ECC9, en donde las personas dominan solamente determinar frecuencias dentro de un conjunto de datos o extraer información a partir de representaciones gráficas. Después están ubicadas en el ECC 5 las personas que dominan estos dos. El estado conglomerado de conocimiento en el que se ubican más personas es aquel en el que se dominan todos los atributos, a excepción de aquellos relacionados con secuencias, patrones, identificación de características comunes e identificar el elemento mínimo y/o máximo de un conjunto de datos.

Según Artavia-Medrano (2015), se puede establecer trayectorias de aprendizaje para obtener información útil para definir la manera en que se puede mejorar el dominio de las habilidades de manera progresiva. Por ejemplo, la trayectoria 1 indicaría el camino que debe seguir alguien que inicialmente domina el conteo (AT7), después se le puede enseñar a extraer información a partir de representaciones gráficas (AT5) y así sucesivamente hasta que domine todos los atributos, incluyendo aquellos en los que se evidenció que esta muestra no alcanzó a dominar por completo.

Tabla 28

Trayectorias de aprendizaje según estados conglomerados de conocimiento.

Trayectoria	**Proceso de la trayectoria**
1	ECC8 → ECC5 → ECC7 → ECC6 → ECC1
2	ECC9 → ECC5 → ECC7 → ECC6 → ECC1
3	ECC9 → ECC2 → ECC3 → ECC4

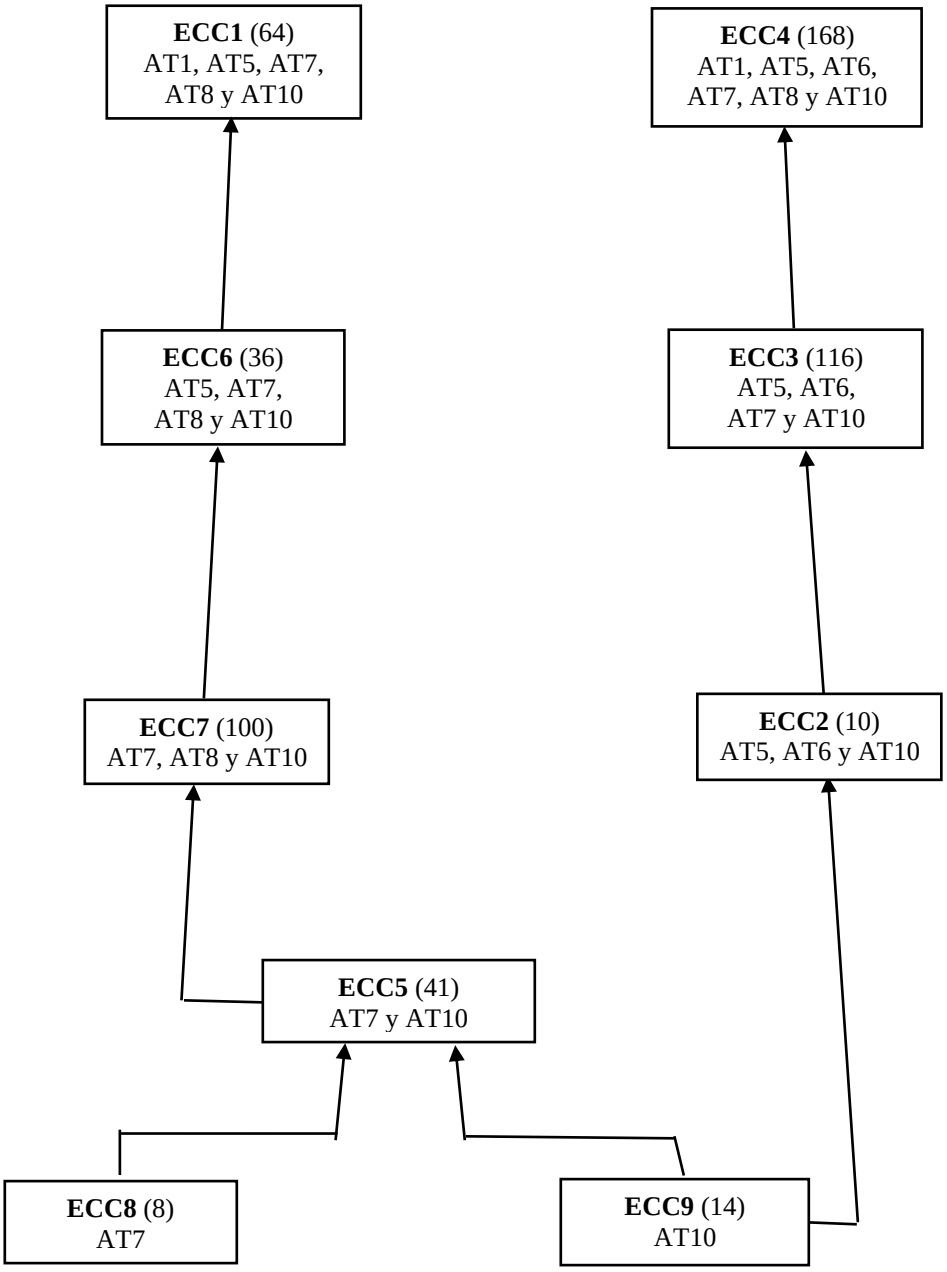

Figura 12. Red de Estados Conglomerados de Conocimiento.

Conclusiones

La aplicación exitosa de un Modelo de Diagnóstico Cognitivo como el *Rule Space* en la implementación de una Evaluación de Diagnóstico Cognitivo en el contexto colombiano hace que este sea un estudio pionero en el país. No existen antecedentes de la implementación de una EDC, que incluya la elaboración de una prueba para ello, y menos con resultados tan útiles, teniendo en cuenta que el contenido evaluado y los resultados de la evaluación pueden relacionarse directamente con aquello que los estudiantes aprenden en el aula. Además de lo novedoso y lo útil de este estudio, hay que mencionar su relevancia: evaluar Pensamiento Estadístico es importante porque se relaciona con el pensamiento crítico, la capacidad para decidir en contextos democráticos e incluso en decisiones concernientes a la salud de las personas. Conocer las dificultades y fortalezas de los estudiantes colombianos en Pensamiento Estadístico puede impulsar transformaciones curriculares y pedagógicas que desemboquen en la mejora en el desempeño académico, pero también en el desarrollo de su vida diaria. Finalmente, evaluarlo en etapas escolares tempranas puede tener un mayor impacto debido a que se pueden prevenir los errores o los vacíos que muestran los estudiantes más adelante.

Por otra parte, la prueba de Pensamiento Estadístico desarrollada para este trabajo es el resultado de un proceso sistemático, apoyado por docentes expertos y fundamentado por las experiencias acumuladas del Laboratorio de Psicometría de la Universidad Nacional de Colombia. Las especificaciones de la prueba fueron planteadas en detalle y responden de

manera fiel al contexto en el cual sería utilizada. Entre las evidencias preliminares de validez que se encontraron hay por lo menos cuatro:

1. Basadas en el contenido de la prueba: Sireci y Faulkner-Bond (2014) indican que hay cuatro componentes que asegurarían la "validez de contenido". El primero es elaborar una definición del dominio, que se logró usando como base los estándares y la validación de los expertos. El segundo es la representación del dominio y el tercero la relevancia, es decir, la adecuación entre la definición del dominio y lo que evalúa el test en su conjunto sumado a la importancia de cada ítem en la representatividad del dominio evaluado, esto se logró debido a que cada componente de la prueba se evaluó por lo menos dos veces y se aseguró que todos los componentes estuvieran representados en la prueba. El último componente hace referencia a la calidad técnica el proceso del desarrollo de la prueba, lo cual es, precisamente, una de las fortalezas más importantes de este estudio.
2. Basadas en la relación con otras variables: El desempeño en la prueba está relacionado con el grado de los estudiantes. En la sección de resultados de la primera fase, se observó que existían diferencias significativas en el desempeño en la prueba entre estudiantes de segundo y tercero (aquellos que están en proceso de dominar los componentes propuestos) y los estudiantes de cuarto y quinto (aquellos que ya "deberían dominar los componentes"), lo que da cuenta de que la prueba es sensible al proceso de enseñanza. También se observaron diferencias en el desempeño de estudiantes según la ciudad y el tipo de colegio (Oficial vs No Oficial), lo que muestra que los resultados de la prueba se dan de acuerdo a las

dinámicas educativas colombianas, donde los colegios No Oficiales muestran en su conjunto una calidad más elevada en relación con los colegios Oficiales, y donde ciudades históricamente relegadas a los últimos lugares en calidad educativa quedan rezagadas también en la prueba de Pensamiento Estadístico. Finalmente, el hecho de que no se encuentren diferencias en el desempeño según el género indica que la prueba es fiel a lo definido en los estándares y al parecer los resultados dependen casi exclusivamente del proceso educativo.

3. Basadas en el proceso de respuesta: El hecho de haber utilizado ítems de respuesta construida fue una ventaja para poder obtener respuestas genuinas por parte de los estudiantes, teniendo en cuenta que debían elaborar la respuesta por sí mismos. La prueba tuvo dos tercios de preguntas de este tipo, por lo que la información sobre el desempeño de los estudiantes tiene niveles mínimos de "adivinación". Asimismo, de las respuestas se podría extraer información importante sobre la forma en que los estudiantes procesan la información y elaboran sus respuestas. Esto puede ser interesante para futuras investigaciones.
4. Basadas en las consecuencias de la evaluación: La prueba de Pensamiento Estadístico tiene evidencias de validez de acuerdo a la gran utilidad que pueden tener sus resultados para mejorar el proceso educativo de los escolares. También puede tener consecuencias positivas si los estudiantes logran dominar el Pensamiento Estadístico y esto se extrapola a su vida cotidiana. Si la información extraída en la evaluación genera cambios curriculares o en las prácticas dentro del aula, las consecuencias podrían tener un mayor impacto positivo.

Evaluación de Pensamiento Estadístico con *Rule Space*

Respecto a la aplicación del método *Rule Space,* se obtuvieron resultados satisfactorios. En este estudio se utilizaron los índices de adecuación de la matriz Q tradicionales en el modelo y el hecho de tener una tasa de clasificación tan alta, da cuenta de que los datos se ajustaron adecuadamente. La utilización de *Rule Space* mostró ser una opción viable para la implementación de Evaluaciones de Diagnóstico Cognitivo en nuestro contexto, ya que se trata de un modelo con un uso simple y de fácil interpretación, pero que entrega información muy relevante para fines educativos. Otro beneficio de la implementación del *Rule Space* tiene que ver con la necesidad de adoptar un modelo cognitivo o teoría, en este caso sobre el pensamiento estadístico; esto provoca debate sobre los constructos medidos (su definición, estructura, componentes, coherencia, etc.), colocando sobre la mesa la necesidad de contar con fundamentación teórica o empírica.

La implementación del *Rule Space* tuvo la limitación de que la base para definir los componentes por evaluar fue débil. Los estándares del MEN no se encuentran adecuadamente enmarcados en una teoría y no existe un programa de investigación que controvierta o demuestre la estructura propuesta allí para el componente de matemáticas. Sin embargo, lo trabajado en este estudio puede plantear la necesidad de que se comience a evaluar su pertinencia. Una limitación adicional, pero también otra oportunidad, es la falta de consenso dentro del tema de la teorización sobre el pensamiento estadístico. Muchos autores han trabajado el tema, pero en este punto se está lejos de establecer acuerdos sobre temas importantes como la definición, componentes, etapas de desarrollo del pensamiento estadístico e incluso sobre su relación con las matemáticas. Las falencias respecto a la definición del tema de pensamiento estadístico y las relacionadas con el sustento teórico y empírico de los

estándares hacen que los atributos propuestos y la jerarquía de los Estados de Conocimiento Conglomerados sean propuestas plausibles, listas para controvertir o demostrar.

La dificultad para encontrar un consenso sobre la definición del pensamiento estadístico y sus componentes demuestra que se trata de un dominio complejo, que además de requerir las habilidades planteadas en los EBC requiere habilidades básicas como lectura, escritura y comunicación. Por eso se debió incluir en la lista de atributos aquí propuesta la habilidad de "Interpretar situaciones descritas en el enunciado". Por ejemplo, esto es coherentes con las propuestas de Gal (2002) sobre los componentes del Pensamiento Estadístico en adultos, donde indica que las habilidades comunicacionales básicas son necesarias para el desarrollo de este tipo de pensamiento. Las habilidades de lectura y escritura facilitan la aprehensión de información contextual, lo que es determinante para poder desarrollar las otras habilidades que hacen parte del Pensamiento Estadístico, esto es coherente con la propuesta de Watson y Callingham (2005) que identifican como factor relevante la forma en que los estudiantes se relacionan con el contexto. En esta investigación se plantea la importancia de habilidades que, por lo menos en los EBC, no se relacionan explícitamente con el Pensamiento Estadístico, pero que se deben tener en cuenta y así propiciar investigación al respecto.

Otra limitación de este trabajo está relacionada con la retroalimentación al desempeño de aquellas personas que no pudieron ser clasificadas en algún EC. Si bien *Rule Space* cuenta con herramientas para evaluar índices específicos de ajuste de las personas que no se pudieron clasificar, como el Índice de Conformidad con la Norma (en inglés *Norm Conformity Index-NCI*), en este estudio no se llevaron a cabo. Es importante que se tenga en cuenta estas

personas ya que se puede encontrar información interesante, a partir de esos patrones de respuesta inesperados y tal vez sean estos estudiantes los que más requieran un diagnóstico de su desempeño.

Este estudio significa una contribución importante para la línea de investigación Métodos e Instrumentos de Investigación en Ciencias del Comportamiento, al incursionar en el campo de las Evaluaciones de Diagnóstico Cognitivo y sumando un estudio más a la larga lista de aplicaciones que buscan mejorar la calidad de la evaluación en psicología y educación en Colombia. Este trabajo deja abierta la puerta para desarrollar nuevos trabajos: además de lo que se mencionó arriba sobre la posibilidad de probar la teoría propuesta sobre el Pensamiento Estadístico y controvertir la propuesta de los estándares, queda pendiente una completa agenda relacionada con la ampliación de las aplicaciones del *Rule Space,* la comparación con otros modelos, su uso en la evaluación de otros dominios y en otros niveles escolares, así como el trabajo para mejorar el acceso al uso de este modelo.

Otra investigación futura que es interesante tiene que ver con el examen a nivel cognitivo de las fuentes de dificultad de los atributos identificados. Es decir, encontrar explicaciones fundamentadas sobre este hallazgo. También es posible que en un futuro se trate de explicar la jerarquía encontrada en los Estados Conglomerados de Conocimientos y relacionarlo con las demandas cognitivas en cada uno de ellos.

Finalmente, es pertinente recalcar que este trabajo plantea la oportunidad de mejorar los procesos de evaluación masiva en Colombia, llegando más allá de la división entre su función sumativa y formativa. El diagnóstico cognitivo es relevante porque entrega información

detallada a todo el sistema (estudiantes, docentes, padres de familia, directivos, investigadores, etc.) y si es lo bastante reveladora puede desencadenar cambios en toda la escuela. Si se adopta el uso de las Evaluaciones de Diagnóstico Cognitivo, los resultados de las evaluaciones masivas serán herramientas útiles, combinando la información que ya entregan las pruebas, pero complementándola con un diagnóstico preciso y fundamentado y así cumplir con la función transformadora que debe tener le educación en nuestra sociedad.

Referencias

Aizikovitsh-Udi, E., Kuntze, S. & Clarke, D. (2016). Connections Between Statistical Thinking and Critical Thinking: A Case Study. En Ben-Zvi, D. & Makar, K. (Eds.), The Teaching and Learning of Statistics: International Perspectives. Springer.

American Educational Research Association, American Psychological Association & National Council of Masurement in Education (2014). Standards for Educational and Psychological Testing. Washington, AERA.

Arnold, F. (2017). Statistical Literacy in Public Debate – Examples from the UK 2015 General Election. *Statistics Educaion Research Journal 16(1), 217-227.*

Artavia-Medrano, A. (2015). Interpretación y análisis de pruebas educativas y psicológicas con el método rule space. *Actualidades en Psicología 29(119) 63-77.*

Artavia-Medrano, A. (2014). Evaluación cognitiva diagnóstica en Matemática: modelo elaborado con el método rule space para estudiantes costarricenses de undécimo año. (Tesis doctoral inédita). Universidad de Costa Rica, Costa Rica.

Artavia-Medrano, A. & Larreamendy-Joerns, J. (2012). Información cognitiva a partir de pruebas de gran escala: El método de representación del espacio de reglas. *Universitas Psychologica 11(2) 599-610.*

Ben-Zvi, D. & Makar, K. (2016). International Perspectives on the Teaching and Learning of Statistics. En Ben-Zvi, D. & Makar, K. (Eds.), The Teaching and Learning of Statistics: International Perspectives. Springer.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71 (3), 425-440.*

Buck, G., Tatsuoka, K. & Kostin, I. (1997). The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension. *Language Learning, 47: 423-466.* doi:10.1111/0023-8333.00016

Chen, Y. (2006). Cognitively diagnostic examination of Taiwanese mathematics achievement on TIMSS-1999. Disertación doctoral no publicada, Arizona State University, Arizona, EE. UU.

Chen, Y., Gorin, J., Thompson, M. & Tatsuoka, K. (2006, abril). Verification of cognitive attributes required to solve the TIMSS-1999 mathematics items for Taiwanese students. Documento presentado en la reunión de la American Educational Research Association, San Francisco, CA, EE. UU.

Cox, D. R., & Efron, B. (2017). Statistical thinking for 21st century scientists. Science advances, 3(6), e1700768. doi:10.1126/sciadv.1700768

Cui, Y., Gierl, M. & Guo, Q. (2016). *The Rule Space and Attribute Hierarchy Methods*. En J.P. Leighton y M.J. Gierl (Eds.), The Handbook of Cognition and Assessment: Frameworks, Methodologies and Applications. Willey and Sons.

Dean, M. (2006). *Item attributes for explaining TIMSS advanced mathematics test performance*. Tesis Doctoral, Universidad de Columbia. EE.UU.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76, 179-199.*

de la Torre, J. & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa, 20, 89-97.*

DiBello, L., Roussoss, L. & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics, 26, 1-50.* DOI: 10.1016/S0169-7161(06)26031-0

Dogan, E. & Tatsuoka, K. (2008). An International Comparison Using a Diagnostic Testing Model: Turkish Students' Profile of Mathematical Skills on TIMSS-R. *Educational Studies in Mathematics, 68 (3) 263-272.*

Dogan, E. (2006). Establishing construct validity of the mathematics subtest of the university entrance examination in Turkey: A rule space application. Disertación doctoral no publicada, Teachers College, Columbia University, New York, EE. UU.

Engel, J. (2017). Statistical Literacy for Active Citizenship: A Call for Data Science Education. *Statistics Educaion Research Journal 16(1), 44-49.*

Gal, I. (2002). Adults' statistical literacy: Meaning, components, responsibilities. *International Statistical Review, 70 (1), 1–25.*

García, P. E., Olea, J. & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26 (3), 372-377.*

Garfield, J. (2011). *Statistical Literacy, Reasoning, and Thinking*. En: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin.

Garfield, J. & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review, 75(3), 372-396.*

Gierl, M. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44 (4), 325-340).*

Gierl, M., Leighton, J. & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19(3), 34-44.*

González, H. & Kuenzi, J. (2012) Science, Technology, Engineering, and Mathematics (STEM) Education: A Primer. Congressional Research Service. Recuperado el 5 de junio de 2015 de https://fas.org/sgp/crs/misc/R42642.pdf

Guerrero, A. (2001). Cognitively diagnostic perspectives on English and Spanish test of mathematics aptitudes. *Dissertation Abstracts International, 62*(08), 543B. (UMI No. 3005725).

Huff, K. & Goodman, D. (2007). *The Demand for Cognitive Diagnostic Assessment*. En J.P. Leighton y M.J. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications. Cambridge: Cambridge University Press.

ICFES (2013). COLOMBIA EN PISA 2012: Informe nacional de resultados, resumen ejecutivo. Recuperado de http://www2.icfes.gov.co/resumen-ejecutivo-de-los-resultados-de-colombia-en-pisa-2012

ICFES (2015). Resumen Ejecutivo Colombia en PISA 2015. Recuperado de http://www.icfes.gov.co/docman/institucional/home/2785-informe-resumen-ejecutivo-colombia-en-pisa-2015/file

Katz, I., Martinez, M., Sheehan, K. & Tatsuoka, K. (1993). Extending the rule space model to a semantically-rich domain: Diagnostic assessment in architecture (Technical Report RR-93-42-ON). Princeton, NJ: Educational Testing Service.

Kuntze, S., Aizikovitsh-Udi, E. & Clarke, D. (2017). Hybrid task design: connecting learning opportunities related to critical thinking and statistical thinking. *ZDM Mathematics Education. 49, 923-935.*

Leighton, J., & Gierl, M. (2007). *Why cognitive diagnostic assessment?* En J.P. Leighton y M.J. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications. Cambridge: Cambridge University Press.

Li, X. & Wang W. C. (2015). Assessment of Differential Item Functioning Under Cognitive Diagnosis Models: The DINA Model Example. *Journal of Educational Measurement, 52 (1), 28–54.*

Liu, J., Ying, Z. & Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika, 80 (2), 468-490.* DOI: 10.1007/S11336-013-9395-4

MEN (2006). Estándares básicos de competencias en lenguaje, matemáticas, ciencias y ciudadanas: Guía sobre lo que los estudiantes deben saber y saber hacer con lo que aprenden. Recuperado el 25 de noviembre de 2015 de http://www.mineducacion.gov.co/1621/articles-116042_archivo_pdf.pdf

Moore, D. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review, 65, 123-165.*

Newell, A. & Simon, H. A. (1972). *Human problem solving. Englewood Cliffs,* NJ: Prentice Hall.

Nichols, P. D. (1994). A Framework for Developing Cognitively Diagnostic Assessments. *Review of Educational Research, 64(4), 575–603*. https://doi.org/10.3102/00346543064004575

Norris, S. P., Macnab, J. S. & Phillips, L. M. (2007). *Cognitive Modeling of performance on Diagnostic Achievement Tests.* En J.P. Leighton y M.J. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications. Cambridge: Cambridge University Press.

Pellegrino, J. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa. 20, 65-77.*

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press

Rupp, A., Templin J., & Henson, R. (2010). *Diagnostic measurement. Theory, methods, and applications*. New York, NY: The Guilford Press.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Stout, W. (2002). Psychometrics: from practice to theory and back. *Psychometrika, 67 (4), 485-518.*

Svetina, D., Gorin, J. & Tatsuoka, K. (2011). Defining and Comparing the Reading Comprehension Construct: A Cognitive-Psychometric Modeling Approach. *International Journal of Testing. 11. 1-23.*

Tarr, J. E., & Lannin, J. K. (2005). *How can teachers build notions of conditional probability and independence?* En G. A. Jones (Ed.), Exploring probability in school: Challenges for teaching and learning (pp.216-238). Nueva York: Springer.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. En N. Frederiksen, R. Glaser, A. Lesgold & M. Shafto (Eds.), Diagnostic monitoring of skills and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. (1991). *Boolean algebra applied to determination of the universal set of misconception states.* (ONR- Reporte Técnico RR-91-44) Educational Testing Service, NJ, Princeton, EE.UU.

Tatsuoka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20(4), 345-354.*

Tatsuoka, K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge Taylor & Francis Group.

Tatsuoka, K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*(4), 901–926.

Tatsuoka K., & Tatsuoka, C. (2005). Stability of classification results on the cognitive diagnosis for individuals. *Japanese Journal for Research on Testing*, *1*(1).

Tatsuoka, K., Varadi, F. & Tatsuoka, C. (2004). PMAIN - RULE SPACE. [Software no publicado]. Trenton, NJ, EE.UU.: Tanar.

Tong, C. (2019). Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. *The American Statistician, 73, (supp 1), 246-261.*

UNESCO (2015). Economic growth and test scores on maths. Recuperado el 30 de mayo de 2015 de http://www.ibe.unesco.org/en/themes/quality-systemic-approaches/quality-framework-geqaf/technical-notes/economic-growth-and-test-scores-on-math.html

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61, 287–308.*

Watson, J. M. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2(2), 3–46.*

Watson, J. M. & Callingham, R. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education.* International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 2004 (pp. 116–162). Voorburg, The Netherlands: International Statistical Institute.

Wegwarth O. & Gigerenzer G. (2018) The Barrier to Informed Choice in Cancer Screening: Statistical Illiteracy in Physicians and Patients. In: Goerling U., Mehnert A. (eds) Psycho-Oncology. Recent Results in Cancer Research, vol 210. Springer, Cham.

White House Office of Science and Technology Policy (2014). Preparing Americans with 21st Century Skills Science, Technology, Engineering, and Mathematics (STEM) Education in the 2015 Budget. Recuperado el 6 de junio de 2015 de https://www.whitehouse.gov/sites/default/files/microsites/ostp/Fy%202015%20STEM%20ed.pdf

Wild, C. J. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67(3), 223-248.*

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8, 370-371.*

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education, 16(2).*

Lista de Tablas

Lista de Figuras

Lista de Anexos