Estudio de validez del examen de Estado *Saber 11* de inglés

Study of the validity of *Saber 11* English exam

Alexis A. López[1]
Jhonny Ropero Pacheco[2]
Julio César Peralta[3]

Resumen

En el año 2005, el Ministerio de Educación Nacional - MEN de Colombia desarrolló un programa a nivel nacional para mejorar y fortalecer la enseñanza y aprendizaje del inglés. Como una de las primeras iniciativas de este proyecto, el MEN diseñó dos exámenes de inglés alineados con el Marco Común de Referencia Europeo. Este manuscrito presenta los hallazgos de un estudio que evaluó la validez de uno de esos exámenes, el examen de Estado *Saber 11 de inglés*. Los datos para este estudio se recogieron a partir de sesiones de evaluación de contenido con los profesores, los resultados del examen, algunos protocolos de *pensar en voz alta* con los estudiantes y una encuesta a estos últimos. Nuestros hallazgos sugieren que los resultados de este examen no son válidos para medir las habilidades lingüísticas generales en inglés de los estudiantes de grado 11 en Colombia

Palabras clave

Validez, utilidad de una evaluación, validación, evidencia de validez.

Abstract

In 2005, the Colombian National Ministry of Education developed a nationwide program to improve and strengthen the teaching and learning of English. As one of this project's initial steps, the Ministry developed two English tests aligned to the Common European Framework of Reference (CEFR). This paper presents the findings of a study that examined the validity of one of these tests, the *Saber 11 English Exam*. Data for this study was gathered using content evaluation sessions with teachers, results from the test, think-aloud protocols with students, and a student survey. Our findings suggest that this test may not be a valid measure of the students' general English language proficiency.

Key words

Test validity, test usefulness, validation, validity evidence.

Artículo recibido el 28 de marzo de 2011 y aprobado el 19 de agosto de 2011

1 Profesor Universidad de los Andes. Correo electrónico: allopez@uniandes.edu.co

2 Profesor Universidad Pedagógica Nacional. Correo electrónico: jropero@pedagogica.edu.co

3 Profesor Universidad de los Andes. Correo electrónico: jucepeli77@gmail.com

Introducción

Los documentos sobre el Programa Nacional de Bilingüismo 2004-2019 (más conocida como Colombia Bilingüe) racionalizan el bilingüismo en Colombia en dos lenguas –español-inglés– como una reacción resultante de las presiones económicas mundiales: en tiempos de globalización, el país necesita desarrollar la capacidad de sus ciudadanos para poder expresarse en una lengua extranjera. En este contexto, Colombia Bilingüe propone nuevas normas de la competencia comunicativa en lengua extranjera: inglés (Ministerio de Educación Nacional, 2005). Este Programa Nacional de Bilingüismo y la inclusión del Marco Común Europeo de Referencia (MCER) alineado a la enseñanza y a los exámenes que son diseñados, pueden traer muchas cosas positivas en el sistema educativo de Colombia. El ICFES (Instituto Colombiano para la Evaluación de la Educación) escribe que el uso del MCER actuará como una fuente de información en la construcción de indicadores de evaluación en el servicio del sector educativo, a fin de fomentar la evaluación de los procesos institucionales, la formulación de políticas y facilitar la toma de decisiones en todos los niveles del sistema educativo (ICFES, 2009: párr. 3).

Para medir los posibles cambios en el dominio del inglés, derivados del programa Colombia Bilingüe, el Ministerio Nacional de Educación - MEN incluyó en su mandato de 2005 normas de medición basadas en el MCER. Según este plan, en 2019 todos los estudiantes graduados de escuelas secundarias deben estar en el nivel B1, mientras que los estudiantes graduados universitarios lo harán en un nivel B2 de acuerdo con el MCER (Ministerio de Educación Nacional, 2005).

Para medir si los graduados de la escuela secundaria han cumplido con los dos objetivos citados en Colombia Bilingüe, el MEN también ha encomendado el diseño y la aplicación de dos exámenes de competencias basados en el MCER (Ministerio de Educación Nacional, 2005). Bajo la dirección del ICFES y en colaboración con el British Council-Colombia y Cambridge ESOL, se diseñaron los exámenes de inglés de *Saber 11* y *Saber Pro*. El examen de Estado *Saber 11 de inglés* está diseñado para ser tomado por todos los estudiantes de grado 11 como parte de una serie de exámenes administrados durante el último año de la escuela secundaria y el otro examen está diseñado para ser tomado por estudiantes de último año de programas de pregrado.

El examen de Estado saber 11 de inglés

Los estudiantes toman el examen de Estado *Saber 11* durante su último año escolar de secundaria. Éste, en su sección de inglés, está dividido en siete partes, todas ellas con preguntas de selección múltiple (45 preguntas). El examen sólo evalúa las siguientes habilidades: lectura, vocabulario y gramática. Competencias como escuchar, hablar y escribir no son evaluados por razones de practicidad (dificultad en la creación de tecnología adecuada para evaluar la habilidad de escucha y dificultad para la formación de evaluadores de habla y escritura). Una muestra completa del examen de Estado *Saber 11 de inglés* se puede descargar desde la página web del ICFES. A continuación ofrecemos una breve descripción de cada parte de la prueba.

Parte 1. Esta parte incluye cinco preguntas de selección múltiple con tres opciones cada una. Se evalúa la capacidad de los estudiantes para entender avisos y señales. La tarea requiere que los estudiantes lean el aviso o señal y luego identifiquen dónde sería visto al elegir la mejor opción.

Parte 2. Esta parte incluye cinco preguntas que se deben relacionar con ocho opciones. En ella se pide a los estudiantes hacer coincidir los elementos de una columna con las definiciones de una lista de opciones de una categoría léxica (por ejemplo, cosas que puedes encontrar en una cocina, profesiones y materiales de aula). Las preguntas son definiciones similares a las que se encuentran en un diccionario.

Parte 3. Esta parte incluye cinco ítems de opción múltiple con tres opciones cada uno, donde se solicita al estudiante terminar una conversación. Todas las preguntas son conversaciones cortas entre dos interlocutores; en la primera parte se encuentra el enunciado

de una de las personas (lo que dice) y las opciones son tres respuestas posibles. Se evalúa la capacidad de comprender y utilizar el idioma inglés en la cotidianidad.

Parte 4. Esta parte está orientada hacia la gramática, compuesta por ocho ítems con tres opciones de respuesta cada uno. Se debe leer un texto informativo corto. En este hay algunas palabras que faltan y los estudiantes deben elegir la mejor opción para completar cada uno de los espacios. Las palabras pueden ser formas verbales, artículos, preposiciones, conjunciones, pronombres, entre otros.

Parte 5. Esta parte se centra en la comprensión de un texto escrito, y al igual que las otras, cuenta con preguntas de selección múltiple, con tres opciones. Esta parte requiere que se lea un breve texto auténtico modificado y que a continuación se respondan algunas preguntas relacionadas con información relevante del texto.

Parte 6. Esta parte consta de cinco preguntas, con cuatro opciones cada una. El estudiante debe realizar una lectura para responder a preguntas de comprensión basadas en un texto breve, modificado pero auténtico. Las preguntas están relacionadas con el propósito del autor, la actitud u opinión, y significados inferenciales, recordando los detalles y el significado global.

Parte 7. Esta parte tiene diez preguntas de cuatro opciones de selección múltiple cada una. Se evalúa tanto la gramática como el vocabulario. A los estudiantes se les pide leer un texto corto informativo. En el texto hay algunas palabras que faltan y los estudiantes tienen que elegir la opción que mejor encaje en cada espacio en blanco. Las preguntas de gramática son palabras de función y las preguntas de vocabulario son palabras de contenido.

Marco teórico

Validez

El concepto de validez ha cambiado a lo largo de los años. Inicialmente, ésta era considerada como una característica de una evaluación, pero ahora la validez es vista como un argumento relacionado con cómo se utilizan y se interpretan los resultados de una evaluación (Chapelle, 1999). Tradicionalmente, la validez se definía como «el grado en que un examen mide lo que dice, o lo que pretende medir» (Brown, 1996, p. 231). Seguido, la validez se subdividía generalmente en tres categorías diferentes: validez de contenido, validez de criterio y validez de constructo (Bachman, 1990). La validez de contenido se refiere a la medida en que una prueba refleja el dominio de contenido que se pretende medir; la validez de criterio se refiere a la medida en que una prueba podría ser utilizada para sacar conclusiones sobre el criterio, y la validez de constructo se refiere a la medida en la que una prueba mide un rasgo psicológico o concepto teórico (Bachman, 1990). Messick sostiene que «las inferencias relacionadas con el contenido son inseparables de las inferencias relacionadas de constructos» (1988, p. 38).

Messick (1989) define la validez como un criterio general de evaluación del grado en el que la evidencia empírica y los fundamentos teóricos dan apoyo a la adecuación y pertinencia de las interpretaciones de los resultados de los exámenes y de las acciones realizadas. En este marco, Messick presenta un concepto unificado de validez y en vez de hablar de diferentes tipos de validez, él habla de diferentes tipos de evidencias. De acuerdo con su modelo de validez, hay dos principales amenazas en contra de la validez de una evaluación: el constructo no está bien representado y la presencia de constructos irrelevantes. El constructo no está bien representado cuando es demasiado estrecho y no incluye aspectos importantes del mismo. Por otra parte, la presencia de constructos irrelevantes ocurre cuando este es demasiado amplio y contiene variables extrañas que podrían hacer la evaluación innecesariamente difícil (dificultad irrelevante del constructo) o demasiado fácil (facilidad irrelevante del constructo) para algunos estudiantes.

Hoy en día, el proceso de validación involucra la acumulación de evidencias para proporcionar una base científica sólida para la interpretación de los resultados propuestos. La evidencia se puede basar

en el contenido del examen, los procesos de respuesta, la estructura interna, las relaciones con otras variables y las consecuencias de las pruebas (AERA, APA y SNEM, 1999). Por el contenido de la prueba, nos referimos a la especificación de los límites del dominio del constructo que está siendo evaluado, los conocimientos, habilidades o competencias que se revelan por cada tarea de evaluación (Messick, 1989). Los procesos de respuesta se refieren a la correspondencia entre los procesos cognitivos que los examinados realmente utilizan al completar una acción y los procesos que se deben usar (AERA, APA y SNEM, 1999). La estructura interna de una prueba es la forma en que sus diferentes partes están relacionados entre sí (AERA, APA y SNEM, 1999). No obstante, las asociaciones con otras variables se refieren a la coincidencia entre los resultados en una prueba con resultados en pruebas similares (AERA, APA y SNEM, 1999). Por último, las consecuencias de las pruebas se refieren a las consecuencias sociales del uso de una prueba en particular para un fin determinado (Messick, 1989). De hecho, «las pruebas comúnmente se administran a la expectativa de que algún beneficio provenga de los resultados [...] Uno de los propósitos fundamentales de la validación es para indicar si estos beneficios específicos se realizan» (AERA, APA y SNEM, 1999, p. 16). A continuación se describen los criterios de validez presentados por Bachman y Palmer (1996), que son específicos para examinar el lenguaje.

Utilidad de la evaluación

Aunque la mayoría de los debates de validez se hacen todavía con base en el marco tradicional, ha habido propuestas sobre la necesidad de criterios alternativos de validez adicionales, tales como la conceptualización de la utilidad de la evaluación de Bachman y Palmer (1996). Su marco de utilidad de la evaluación incluye la validez de constructo, la confiabilidad, autenticidad, interactividad, funcionalidad e impacto. Bachman y Palmer (1996) definen estas cualidades de la prueba:

- La validez de constructo se refiere a «El significado y las apropiadas interpretaciones que hacemos sobre los resultados de la evaluación» (p. 21).
- La confiabilidad se refiere a la «consistencia de la medición» (p. 19).
- La autenticidad se refiere al grado de correspondencia entre las tareas o preguntas en la evaluación y las características del uso de la lengua objetivo en contextos reales (p. 23).
- Interactividad se refiere a las estrategias y procesos cognitivos que usan los estudiantes para completar las tareas en una evaluación (p. 25).
- Aplicación en la práctica se refiere a «la relación entre los recursos que serán necesarios en el diseño, desarrollo y uso de los mismos en la evaluación, que estarán disponibles para estas actividades» (p. 36).
- El impacto se refiere al efecto que la evaluación tiene en «la sociedad, los sistemas educativos, y sobre los individuos en esos sistemas» (p. 29).

Existen varios tipos de evidencias que deben ser utilizadas para construir un caso sobre la validez de una evaluación. Esta evidencia es utilizada luego para formar un argumento sobre la validez de la evaluación, el cual, en conjunto, presenta un caso a favor o en contra de las inferencias y supuestos potenciales de dicha evaluación (Kane, 1992). Después de presentar todos los diferentes tipos de evidencia a favor o en contra de la validez de una prueba, se llega a una conclusión sobre la validez de la evaluación (Shepard, 1993).

Ha habido un par de estudios realizados en Colombia que han examinado el impacto de este examen (Barletta Manjarrés, 2005; Barletta Manjarrés y Carrascal, 2006), el impacto del examen *Saber Pro de inglés* (Tejada y Castillo, 2010) y la validez del mismo (López y Janssen, 2010), pero no encontramos ningún estudio que examine la validez del examen de Estado *Saber 11 de inglés*. Por lo tanto, creemos que un estudio como el que presentamos aquí puede contribuir a las discusiones académicas en relación con Colombia Bilingüe, los exámenes del MCER en Colombia o el examen de Estado *Saber 11 de inglés*, ya que presentamos evidencia empírica de la validez de este examen. Podría llegar a ser especialmente relevante si esta evaluación se utiliza para informar al público sobre la eficacia del Programa Nacional de Bilingüismo *Colombia Bilingüe*, para evaluar los

programas de inglés de los colegios colombianos y para medir el desempeño de los estudiantes de bachillerato en el idioma inglés. Con el fin de examinar la validez del examen de Estado *Saber 11 de inglés* nos dispusimos a responder a las siguientes preguntas de investigación:

a. ¿En qué medida el examen de Estado *Saber 11 de inglés*, adecuadamente mide los niveles de desempeño definidos por el Marco Común Europeo de Referencia?
b. ¿En qué medida el examen de Estado *Saber 11 de inglés*, adecuadamente mide las habilidades lingüísticas generales en inglés de los estudiantes de último año de escuelas secundarias en Colombia?

Metodología

En esta sección se ofrece una visión general de los métodos utilizados para examinar la validez del examen de Estado *Saber 11 de inglés*.

Participantes

En este estudio participaron cuatro grupos de personas: 1) profesores que hicieron una evaluación de contenido de la prueba; 2) estudiantes que tomaron la prueba; 3) estudiantes que contestaron una encuesta; y 4) estudiantes que participaron en una sesión de pensar en voz alta. Para cada una de estas sesiones intervinieron diferentes actores, ya que el foco del estudio no eran los participantes sino la validez del examen de Estado *Saber 11 de inglés*. Esto nos permitió triangular información sobre la validez de la prueba desde diferentes perspectivas. A continuación se da información sobre cada uno de los grupos de participantes.

Profesores de inglés. Un grupo de 25 profesores de inglés, de grados 10 y 11 de instituciones públicas y privadas, participaron en la evaluación del contenido del examen de Estado *Saber 11 de inglés* (se describe a continuación). Los participantes fueron seleccionados con base en su disponibilidad para participar en el estudio, su familiaridad con el MCER y su experiencia en la enseñanza de inglés en secundaria.

Estudiantes. Un total de 245 estudiantes de grado once de diferentes instituciones educativas participaron en el estudio. Doce de ellos participaron en una sesión de pensar en voz alta, 69 tomaron el examen y 164 completaron una encuesta. El número de participantes para cada uno de estos instrumentos varía dependiendo de la naturaleza de los datos recogidos. La sesión de pensar en voz alta es un instrumento de recolección de datos cualitativos, por lo tanto se busca profundidad en estos datos (pocos participantes). Para evaluar los resultados de la prueba se requieren como mínimo 50 participantes y para la encuesta se requiere un número más alto. Los estudiantes se seleccionaron de seis colegios públicos, cuatro de Bogotá y dos de la costa atlántica. Estos fueron seleccionados con base en los siguientes criterios: estar cursando grado once, y tener disponibilidad y disposición para completar las actividades propuestas. También se seleccionaron por conveniencia, ya que teníamos acceso a estos colegios (lugares donde habían trabajado dos de los investigadores). Una vez tuvimos acceso a los colegios, los estudiantes fueron identificados por sus profesores de inglés como participantes potenciales. Tratamos de conseguir estudiantes con diferentes niveles de desempeño en inglés (alto, intermedio y bajo), según lo definido por los niveles de clase actual.

Recolección de datos

Se utilizaron cuatro instrumentos diferentes para recoger los datos y recopilar información acerca de la validez del examen de Estado *Saber 11 de inglés*: sesiones de evaluación de contenidos, sesiones de pensar en voz alta, resultados del examen y una encuesta a estudiantes. El propósito de estos cuatro instrumentos fue obtener información acerca del constructo y contenido del examen, su uso, su consistencia interna, las percepciones de los profesores y estudiantes acerca del examen y el impacto del mismo; además, estos instrumentos documentaron las estrategias que utilizan los examinados para completar la prueba.

Sesiones de evaluación de contenido. Se realizaron dos sesiones con profesores de inglés de los grados 10 y 11. En la primera sesión participaron 14 profe-

sores y en la segunda 11. En estas sesiones se realizó una evaluación del contenido del examen, referida a «los juicios de expertos sobre la habilidad que las preguntas del examen miden» (Chapelle, 1999, p. 168). Las sesiones duraron aproximadamente tres horas cada una. Todos los participantes recibieron la prueba, un formato para diligenciar (ver Anexo 1) y una lista de descriptores de los niveles del MCER. El propósito de la evaluación del contenido fue pedir a los profesores de inglés su opinión de expertos acerca de las habilidades que las preguntas del examen están midiendo o las habilidades que los estudiantes tienen que utilizar para completar cada tarea, y criticar las preguntas, poniendo de relieve sus fortalezas y limitaciones. A los profesores también se les pidió alinear cada pregunta en el examen con el MCER y los objetivos que se encuentran dentro de su plan de estudios en cada institución. También se les pidió criticar las preguntas del examen y proporcionar sus juicios sobre los constructos evaluados, los textos y las tareas a realizar.

El protocolo de pensamiento en voz alta. Doce estudiantes de distintas instituciones educativas en Bogotá participaron en una sesión de pensamiento en voz alta. En ella se siguieron en general procedimientos similares a los propuestos por Ericsson y Simon (1993). Los protocolos de pensar en voz alta se utilizan comúnmente como un método de recopilación de datos en la resolución de problemas (Stratman y Hamp-Lyons, 1994). El protocolo utilizado en este estudio fue modelado a los estudiantes (ver Anexo 2), quienes tuvieron la oportunidad de practicarlo después. A todos los estudiantes que participaron en la sesión de pensamiento en voz alta se les pidió que narraran en su idioma de preferencia todo lo que hacían para completar el examen. Las sesiones se llevaron a cabo de forma individual en una oficina privada y fueron grabadas en audio en su totalidad, para que pudieran ser transcritas y analizadas más adelante. Inmediatamente después de completar cada parte del examen, realizamos entrevistas retrospectivas con los estudiantes para recoger información sobre sus percepciones y sobre cómo se sintieron completándolas. Cada sesión y entrevista retrospectiva duró aproximadamente 80 minutos.

Debido a que algunos autores han identificado limitaciones con este tipo de protocolo, en el sentido que las sesiones de pensar en voz alta podrían interferir con la tarea de aprendizaje (Shanks y St. John, 1994) o con la capacidad de atención que les impediría poder asistir al aprendizaje (Jourdenais, 1996), se tomaron las siguientes medidas: 1) no se tuvieron en cuenta los resultados en la prueba de estos 12 participantes para hacer análisis estadísticos; y 2) se realizaron las sesiones en español para facilitar la expresión de sus ideas.

Exámenes. Para analizar la confiabilidad y consistencia interna del examen, 69 estudiantes de diversas instituciones educativas de Bogotá lo tomaron en diferentes sesiones, cada una de las cuales duró aproximadamente 100 minutos. Los exámenes fueron calificados otorgando el valor de un punto a cada pregunta. Una muestra completa del examen se puede descargar desde la página web del ICFES.

Encuesta a estudiantes. Se realizó una encuesta en persona para obtener información sobre el impacto del examen en estudiantes del grado 11 de varios colegios colombianos. Su diligenciamiento duró aproximadamente 20 minutos y se aplicó a 164 estudiantes durante septiembre y octubre de 2009. La encuesta tenía una pregunta cerrada tipo Likert y tres preguntas abiertas (ver Anexo 3).

Análisis de datos

Los datos cuantitativos (resultados del examen) se estudiaron pasándolos a una base de datos, donde luego se analizaron usando SPSS. Se usó estadística descriptiva (tablas de frecuencia, media y desviación estándar) para describir los resultados de los estudiantes. Se calculó la confiabilidad del examen usando el Alfa de Cronbach, método estadístico para determinar la confiabilidad de una prueba (Cronbach, 1951). Se escogió este método puesto que requiere una sola administración de la prueba; además, los principales coeficientes de estimación basados en este enfoque son sencillos de computar y están disponibles como opción de análisis en SPSS. También se calcularon los coeficientes de correlación lineal de Pearson entre cada una de las partes del examen usando SPSS para determinar la

consistencia interna del examen, es decir, qué tanto las diferentes partes del examen miden el mismo constructo (AERA, APA y SNEM, 1999). Por otra parte, los datos cualitativos (las preguntas abiertas de las encuestas, las sesiones de pensar en voz alta y los comentarios de las sesiones de evaluación de contenido) se analizaron de la siguiente manera: los autores transcribieron todos los datos; luego se ordenaron y organizaron todos los materiales recogidos. Individualmente, se leyeron todos los datos varias veces y luego se procedió a utilizar la codificación abierta para analizarlos (Miles y Huberman, 1994). Las categorías emergieron directamente de los datos como ideas repetidas o temas. Esta técnica nos permitió utilizar la triangulación para validar nuestra codificación e interpretación de los investigadores. Las categorías que surgieron a partir de los datos fueron *evidencia basada en el contenido y el constructo del examen* y *evidencia basada en el impacto del examen*.

Limitaciones del estudio

A continuación se presentan algunas limitaciones que se evidenciaron en el estudio. A pesar de estas limitaciones, ninguna de ellas invalida las interpretaciones.

- La resistencia de algunos de los participantes para compartir de manera abierta sus percepciones y opiniones sobre el examen.
- La dificultad para aplicar, en algunas ocasiones, los diferentes instrumentos de recolección de datos por el desarrollo de diferentes actividades en los colegios.
- La falta de evidencia externa (resultados en otras pruebas de inglés) para sustentar la validez del examen.

Aspectos éticos

Para garantizar que el estudio fuera ético y responsable, se tuvieron en cuenta los siguientes aspectos:

1. La participación en el estudio fue totalmente voluntaria.
2. Se les informó a todos los participantes sobre el objetivo del estudio con anticipación y se obtuvo consentimiento informado de cada uno de ellos.
3. Todos tenían el derecho de retirarse del estudio en cualquier momento, sin que esto les fuera a ocasionar consecuencias negativas.
4. De igual manera, se garantizó que el estudio tuviera un carácter confidencial. Por lo tanto, se protegió la identidad de las instituciones y de los participantes durante todo el proceso.

Hallazgos y discusión

En esta sección presentamos diferentes tipos de evidencias a favor y en contra, a partir de las cuales ofrecemos un argumento de la validez del examen de Estado *Saber 11 de inglés*. Según Chapelle (1999), un argumento de validez debe presentar e integrar diferentes tipos de evidencias que permiten llegar a una conclusión sobre la validez de las interpretaciones o inferencias que se pueden hacer de los resultados de un examen. Los diferentes tipos de evidencias que presentamos tienen que ver con el contenido y constructo del examen, la confiabilidad y consistencia interna del examen, los procesos cognitivos que usan los estudiantes, y el impacto del examen.

Evidencia basada en el contenido y constructo del examen

Tal como lo resaltamos en la descripción del examen, éste no evalúa habilidades de escritura, escucha o habla. Este es uno de los puntos más problemáticos en el sentido que se hacen inferencias sobre el nivel de competencia lingüística de los estudiantes con base en los niveles del MCER. Como estos hacen referencia a todas las habilidades lingüísticas, los resultados del examen de Estado *Saber 11 de inglés* pueden llevar a interpretaciones erradas sobre las habilidades de los estudiantes en habla, escucha y escritura. El MCER define cinco capacidades lingüísticas que los estudiantes deben adquirir para cada nivel: 1) comprensión de lectura; 2) comprensión auditiva; 3) interacción social; 4) expresión social; y 5) expresión escrita. Según el análisis de contenidos, encontramos que el exa-

men de Estado *Saber 11 de inglés* solamente evalúa algunas capacidades de comprensión de lectura.

Los resultados del análisis de contenido del examen también evidencian que los profesores de inglés perciben que no hay una alineación adecuada entre el examen y los niveles del MCER (ver tabla 1). Estos resultados sugieren que existe una alineación parcial entre el examen de Estado *Saber 11 de inglés* y las capacidades lingüísticas definidas para cada nivel del MCER. Muchos profesores argumentan que el examen solamente da información sobre algunas habilidades lingüísticas. Una profesora comentó que por ejemplo la Parte 2 del examen evalúa el vocabulario, pero solamente se enfoca en la definición de palabras. El examen no mide la habilidad que tiene el estudiante "de usar o producir esas palabras". Esto implica que los resultados solamente dan información sobre algunos aspectos relacionados con los niveles propuestos por el MCER.

Tabla 1. Percepción de la alineación entre el examen y el MCER.

Nivel de Alineación	Frecuencia
Totalmente alineado	0
Bastante alineado	3
Parcialmente alineado	14
Poco alineado	8
Nada alineado	0

Nota: muestra de 25 profesores

De igual manera, la sesión de análisis de contenido da información sobre la alineación entre el examen y los procesos de enseñanza de inglés como lengua extranjera en colegios de secundaria (ver tabla 2). Según los profesores de inglés, existe una alineación adecuada en el sentido que los tipos de tareas que ellos hacen en clase son parecidos a las actividades propuestas en el examen. Algunos profesores argumentan que ciertas instituciones adoptan políticas de evaluación que requieren que se usen tipos de preguntas similares a las utilizadas en los exámenes de estado. Uno de ellos comentó «El colegio nos obliga a usar preguntas tipo ICFES en los exámenes finales, así que usamos preguntas parecidas en clase para prepararlos a tomar el examen». Otro profesor anotó «siento la responsabilidad de prepararlos para tomar las pruebas ICFES. Por lo tanto uso algunos simulacros en clase y las evaluaciones son similares a las del ICFES». Por otro lado, un docente de un colegio bilingüe expresó que su institución se enfoca en todas las habilidades lingüísticas y presta más atención a los exámenes internacionales que los estudiantes tienen que tomar. Esto sugiere que existe una alineación entre el examen y el proceso de enseñanza más por el poder que tienen los exámenes nacionales y la importancia que le dan las instituciones, los profesores y los estudiantes.

Tabla 2. Percepción de la alineación entre el examen y el proceso de enseñanza.

Nivel de Alineación	Frecuencia
Totalmente alineado	4
Bastante alineado	12
Parcialmente alineado	4
Poco alineado	3
Nada alineado	2

Nota: Muestra de 25 profesores

Otro punto importante que vale la pena resaltar es lo relacionado con la autenticidad del examen. Autenticidad se refiere a qué tan relacionadas son las tareas que tienen que hacer los estudiantes en un examen con las tareas que tendrían que hacer en la vida real (Bachman y Palmer, 1996). Según los docentes que participaron en la evaluación de contenido del examen y la información que se recogió de las sesiones de pensar en voz alta, encontramos que las preguntas en el examen de Estado *Saber 11 de inglés* no son muy auténticas. Un docente comentó que sus estudiantes generalmente leen textos más largos y de diferentes tipos (por ejemplo, cuentos cortos, noticias, etc.).

Los profesores también comentaron que los textos que usan en clase tienden a tener estructuras gramaticales y vocabulario mucho más complejos que las lecturas que aparecen en el examen. Otro profesor comentó que sus estudiantes tienen que usar estrategias cognitivas más complejas cuando leen textos. Por ejemplo, los estudiantes tienen que comparar y contrastar, argumentar, resumir,

establecer relaciones de causa y efecto o reescribir una historia. Los docentes encontraron que los textos que aparecen en las partes cuatro a la siete del examen de Estado *Saber 11 de inglés* no son muy auténticos. Ellos argumentan que estos textos son basados en textos auténticos, pero han sido modificados lingüísticamente, en extensión y en vocabulario, para ajustarse al nivel de suficiencia en inglés de los estudiantes. Los docentes concuerdan que estos textos no se parecen a los textos que ellos leerían en clase o en la vida real (en Internet o en bibliotecas). Uno de ellos explicó que en su clase, «se usan textos mucho más largos de diferentes tipos». En las sesiones de pensar en voz alta, los estudiantes argumentaron que los textos eran muy cortos y que esto facilitaba contestar las preguntas, así fueran de un alto nivel de dificultad. En estas sesiones, igualmente evidenciamos que los estudiantes podían leer el texto varias veces, y como la mayoría de las preguntas eran de tipo inferencial, eventualmente podrían encontrar la información necesaria para contestar una pregunta. Fortus, Coriat y Fund (1998) encontraron que los textos cortos reducen la cantidad de información que tienen que procesar los estudiantes para contestar una pregunta.

Por último, los docentes comentaron que las partes 4 y 7 son las más auténticas del examen, en el sentido que evalúan gramática y vocabulario de una manera contextualizada. Los docentes consideran que esta forma de evaluar es mucho más válida y auténtica que estimar estas habilidades de manera aislada. Pero también resaltan que el tipo de actividad que tienen que hacer (completar frases) no es una actividad muy auténtica. Aunque es una actividad que es común en la enseñanza de una lengua extranjera, no es una actividad que generalmente se haga cotidianamente.

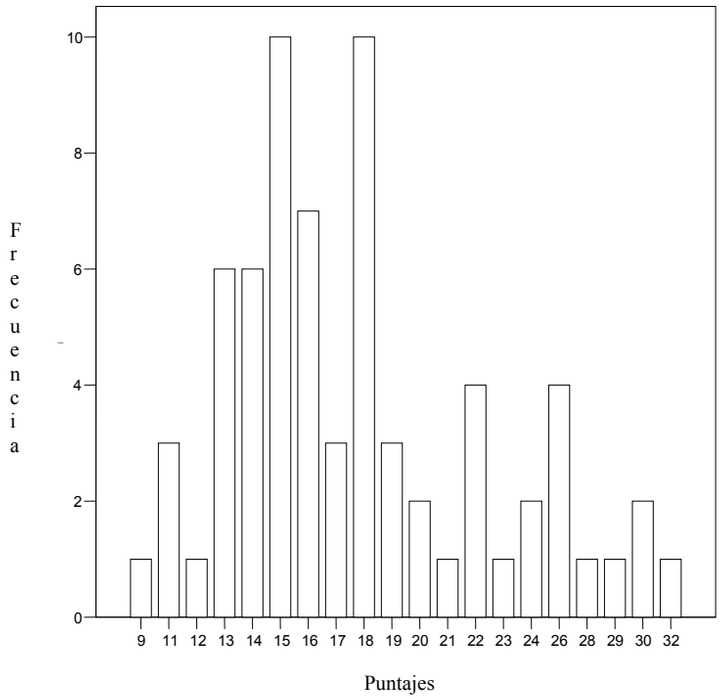
Evidencia basada en la confiabilidad y consistencia interna del examen

En la tabla 3 se muestran los resultados estadísticos descriptivos del examen con base en el desempeño de los 69 estudiantes que lo tomaron. Según estos resultados, observamos que la media (el promedio aritmético) fue de 17.96 sobre un puntaje total de 45 puntos. Esto sugiere que el desempeño de los estudiantes fue bajo y la prueba tiene un nivel de dificultad alto para ellos. Adicionalmente, encontramos que la desviación estándar es de 5.11. Esto sugiere que hay una dispersión alta, lo cual se confirma con un rango alto de 23 puntos. Según la distribución de frecuencias relacionada en la figura 1, vemos que se presenta un sesgo positivo (sesgada a la derecha), lo cual nos lleva a concluir que este examen fue difícil para este grupo de estudiantes de colegios públicos. Esta información se corrobora con los hallazgos de la sesión de pensamiento en voz alta, en donde 10 de los 12 participantes manifestaron que las partes 1 a la 3 son sencillas, pero las últimas cuatro son bastantes difíciles para ellos. El criterio de confiabilidad del examen de Estado *Saber 11 de inglés* se calculó teniendo en cuenta el coeficiente de Alfa Cronbach, el cual determina el grado de consistencia y precisión en la medición de la competencia lingüística en inglés de los estudiantes. Para este estudio, el Alfa de Cronbach fue de 0,665. Este valor está un poco debajo de lo aceptable, ya que se esperan cómo mínimo valores entre 0,70 y 0,90 para inferir que el examen tiene una buena consistencia interna.

Tabla 3. Estadística descriptiva de los resultados del examen.

N	Media	Desviación Estándar	Rango
69	17.96	5.11	23

Figura 1. Distribución de puntajes en el examen



Por otro lado, para examinar la consistencia interna del examen de Estado *Saber 11 de inglés*, se calcularon las intercorrelaciones entre los puntajes en cada una de las siete partes del mismo. El análisis de la estructura interna de un examen da información sobre el grado de relación entre sus diferentes partes y se determina si estas evalúan el mismo constructo. En la tabla 4 se muestran las correlaciones encontradas. Las que están marcadas con asteriscos indican que hay una correlación significativa. Para destacar, se hallaron correlaciones significativas entre las partes 4 y 7 y las partes 5 y 6, sugiriendo que miden constructos similares. En el caso de las partes 4 y 7, miden gramática, y las partes 5 y 6, comprensión de lectura.

Tabla 4. Consistencia interna del examen.

	Parte 1	Parte 2	Parte 3	Parte 4	Parte 5	Parte 6
Parte 2	.140					
Parte 3	.194	.248*				
Parte 4	.171	.262*	.255*			
Parte 5	.022	.223	.240*	.039		
Parte 6	-.028	.280*	.034	.122	.275*	
Parte 7	-.008	.140	.282*	.278*	.162	.090

*La correlación es significativa al nivel 0.05 (bilateral).

Evidencia basada en las estrategias que los estudiantes usan para contestar el examen

En esta sección se presentan evidencias sobre los diferentes tipos de estrategias que utilizan los estudiantes para contestar las preguntas en el examen de Estado *Saber 11 de inglés*. Según los protocolos de pensamiento en voz alta, encontramos que los estudiantes utilizan diversos tipos de estrategias cognitivas para llegar a la respuesta. Por ejemplo, encontramos que muchos estudiantes tratan de adivinar las respuestas o eliminar opciones de la mayoría de las preguntas. Uno de los estudiantes manifestó que «siempre contesto las preguntas así no sepa las respuestas. Lo que hago es tratar de eliminar unas opciones para que sea más fácil adivinar la respuesta». En algunas ocasiones los estudiantes llegaron a la respuesta correcta sin que necesariamente tuvieran el conocimiento o competencia lingüística. El hecho que los estudiantes puedan adivinar correctamente las respuestas representa una amenaza grande a la validez del examen, en el sentido que una respuesta correcta indica qué tanto se conoce del constructo que se está evaluando (Fortus, Coriat y Fund, 1998). De igual manera, se puede considerar adivinar o eliminar opciones como constructo irrelevante (Messick, 1989), en el sentido que hay variables extrañas que hacen que la pregunta sea más fácil de lo que realmente es.

También encontramos que algunas partes del examen miden otras habilidades para llegar a la respuesta correcta. En la parte 1, que debería medir la habilidad del estudiante para leer textos cortos y sencillos, en realidad mide la habilidad de relacionar palabras. En las sesiones de pensar en voz alta, encontramos de manera consistente que todos los estudiantes relacionan una palabra clave del aviso con una de las opciones. Por ejemplo, asociaban *table* (mesa) con *restaurant* (restaurante), o la palabra *basketball* (baloncesto) con *sports center* (centro deportivo), sin que necesariamente entendieran el aviso. Esto nos lleva a concluir que la parte 1 del examen mide más vocabulario que lectura, lo cual nos hace reflexionar sobre su verdadero constructo.

También encontramos que los estudiantes no utilizan bastantes estrategias metacognitivas para contestar las preguntas basadas en textos (partes 4-7); simplemente usan asociación de palabras para determinar las palabras que no conocen o las que hacen falta para completar el texto. No encontramos ninguna evidencia de que los estudiantes usaran otros tipos de estrategias como hacer inferencias, analizar estructuras gramaticales o usar conocimientos previos. Por lo tanto, así el examen mida habilidades de lectura, gramática y vocabulario, solamente se miden algunas competencias relacionadas con dichas habilidades.

Evidencia basada en el impacto del examen

Encontramos que el examen de Estado *Saber 11 de inglés* tiene un impacto en los estudiantes y este se ve reflejado en acciones que toman las instituciones educativas. Por ejemplo, en la encuesta los estudiantes manifestaron que ciertos colegios sí les ofrecen apoyo para prepararse para el examen, aunque algunos piensan que este no es suficiente. Por ejemplo, una estudiante de grado 11 expresó que «el colegio no ayuda mucho, sólo nos brinda vocabulario y algo de pronunciación. Esto no es nada completo para nuestra formación». La mayoría del apoyo que ofrecen las instituciones es con cursos preparatorios, los cuales son más conocidos como *Pre-ICFES*. Otro tipo de preparación que se ofrece, es de usar evaluaciones que tengan la misma clase de preguntas que aparecen en el examen de *Estado Saber 11*. El objetivo principal de muchos de estos apoyos es familiarizar a los educandos con el formato del examen, como lo expresa otro estudiante: «el colegio [nos apoya], en gran parte, ya que la mayoría de los exámenes que realizamos en el colegio son tipo ICFES y esto nos ayuda en gran medida para que cuando lleguemos al examen ICFES y veamos este tipo de preguntas no nos impresionemos».

Otros colegios ofrecen apoyo dentro de los cursos regulares de inglés. Por ejemplo, se hacen actividades de refuerzo, profundizaciones, se aclaran dudas, se hacen simulacros o se usan evaluaciones similares al examen de *Estado Saber 11*. Otros cole-

gios ofrecen cursos de preparación, los cuales generalmente se ofrecen a estudiantes de grado 11, pero en algunas oportunidades se extienden a décimo grado. Estos cursos preparatorios son financiados, en su mayoría, por las Secretarías de Educación, y muchos de ellos se realizan los sábados o una vez por semana, en jornada contraria.

También evidenciamos otro importante impacto del examen de Estado *Saber 11 de inglés* en las acciones tomadas por los profesores de inglés. Según los estudiantes, algunos tratan de implementar elementos del examen en sus cursos regulares. Con base en las respuestas de los estudiantes, podemos evidenciar que los profesores juegan un papel muy importante en su preparación. Por ejemplo, una estudiante comentó que «me ayuda con la experiencia de mi profesor en este campo, dándome una buena capacitación en inglés, tanto en clases como en la ayuda del pre-ICFES, y por último, motivándome a estudiar este idioma tan importante».

Y, por último, encontramos que muchos estudiantes toman acciones por su propia cuenta para prepararse para el examen. Por ejemplo, repasan por su cuenta y realizan actividades como estudiar materiales que se les ha dado, en clase o reforzar lo que han aprendido en el aula. Una estudiante dijo: «leo documentos en inglés, subrayo los verbos o palabras desconocidas y busco su significación en el diccionario. También repaso las clases dadas por el profesor». Otros utilizan Internet para conseguir exámenes parecidos al examen de Estado *Saber 11 de inglés* o buscar materiales para practicar. Otros realizan actividades como ver televisión y escuchar canciones en inglés, estudiar vocabulario y leer textos en dicho idioma. También encontramos que algunos estudiantes toman cursos adicionales para prepararse para tomar el examen, incluyendo cursos pre-ICFES, cursos autodidactas (pe. *Inglés sin barreras*) o cursos virtuales del SENA.

Conclusiones

En este estudio hemos presentado varios tipos de evidencias sobre la validez del examen de Estado *Saber 11 de inglés*. Al analizarlas en su totalidad, podemos crear un argumento sobre la validez del mismo y ver qué tan apropiadas son las inferencias que se pueden hacer, con base en sus resultados, sobre los niveles del MCER y sobre las habilidades lingüísticas generales en inglés de los estudiantes de grado 11 en instituciones colombianas.

Entre las evidencias a favor del examen, encontramos que éste no tiene un impacto negativo grande en los estudiantes, los profesores o en el proceso de enseñanza del inglés. De igual manera, encontramos que la gramática y el vocabulario se evalúan de manera contextualizada, ya que se utilizan lecturas basadas en textos auténticos y no se evalúa cada una de estas habilidades de manera aislada.

Por otro lado, encontramos que existen más evidencias en contra de la validez del examen. La más crítica es que el constructo no está bien representado en el examen. Es decir, el constructo no incluye aspectos relevantes relacionados con habilidades lingüísticas generales en inglés como escritura, habla y escucha, y existen muchos aspectos relacionados con lectura, gramática y vocabulario que no son evaluados. Encontramos también que el examen no está adecuadamente alineado con los niveles del MCER, según los resultados del análisis de contenido que hicieron los profesores. Otro punto en contra es que el lenguaje usado no es auténtico y todas las preguntas son de selección múltiple, lo cual impide, en muchas ocasiones, que los estudiantes usen habilidades cognitivas complejas, y algunas preguntas no permiten que los estudiantes demuestren todas sus habilidades lingüísticas en inglés.

Con base en estas evidencias, concluimos que el examen de Estado *Saber 11 de inglés*, en su versión actual, no es válido para determinar niveles de MCER ni para dar información sobre las habilidades lingüísticas generales en inglés de los estudiantes de grado 11 en Colombia. Tal como se usa e interpreta el examen actualmente, éste determina el nivel de competencia lingüística en inglés con base en el MCER. Por lo tanto, esta clasificación de nivel implica que el estudiante demuestre todas las habilidades lingüísticas descritas en los niveles del MCER, incluyendo habilidades en escritura, habla y escucha. Esto puede llevar a interpretaciones erróneas sobre las habilidades lingüísticas en inglés

de los estudiantes. Consideramos que las únicas inferencias que se pueden hacer con base en este examen, están relacionadas con algunas habilidades de lectura, gramática y vocabulario.

Implicaciones para la evaluación de suficiencia lingüística en inglés

Para poder determinar la efectividad del Programa Nacional de Bilinguismo y las habilidades lingüísticas generales en inglés de los estudiantes de grado 11, es necesario diseñar un examen que dé cuenta de todas las competencias descritas en el MCER. También es importante evaluar las competencias lingüísticas en inglés de los estudiantes durante todo el proceso de enseñanza y aprendizaje, en el aula de clase, y no únicamente al final del proceso, cuando lleguen al grado once. La evaluación debería tener también un carácter formativo, que todos los participantes puedan usarla para mejorar, y no simplemente sumativo, para informar a los estudiantes el nivel que lograron.

Implicaciones para investigaciones futuras

Es importante estudiar la validez de criterio del examen de Estado *Saber 11 de inglés*. Es decir, se deben hacer estudios donde se compare el desempeño de los estudiantes en este examen, con los resultados en otras pruebas similares como el TOEFL, IELTS, PET o KET. También es importante examinar el impacto que tiene este examen en los procesos de enseñanza y aprendizaje del inglés como lengua extranjera en colegios colombianos a la luz del Programa Nacional de Bilingüismo, sobre todo cuando se conoce que este examen no mide habilidades de producción. Es crítico conocer qué tanto define este examen lo que se enseña en el aula de clase. Y por último, sería interesante estudiar qué tanto se puede usar este examen para evaluar la efectividad de los programas de inglés en colegios públicos colombianos y para evaluar el impacto del Programa Nacional de Bilingüismo.

Bibliografía

AERA, APA y SNEM. (1999). Standards for educational and psychological testing. Washington D.C.: American Educational Research Association.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, L. F. y Palmer, A. S. (1996). Language testing in practice: designing and developing useful language tests. Oxford, UK: Oxford University Press.

Barletta, N. (2005). Washback of the foreign language test of the state examination in Colombia: a case study. *Arizona Working Papers in Second Language Acquisition and Teaching*, 12, 1-20.

Barletta, N. y May, O. (2006). Washback of the ICFES exam: a case study of two schools in the Departamento del Atlántico. *Íkala: Revista de Cultura y Lenguaje*, 11, 235-261.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall.

Chapelle, C. A. (1999). Validity in language assessment. *Annual review of applied linguistics*, 19, 254-272.

Ericsson, K. A. y Simon, H. A. (1993). Protocol analysis: verbal reports as data. Cambridge, MA: MIT Press. (2nd ed.).

Fortus, R., Coriat R. y Fund, S. (1998). Prediction of item difficulty in the English Subtest of Israel's Inter-university psychometric entrance test. In A. J. Kunnan (Ed.), *Validation in language assessment* (p.p. 61-87). Mahwah, NJ: Lawrence Erlbaum Associates.

ICFES (2009). Características y guías de ECAES. Recuperado el 3 de marzo de 2009. Disponible en: http://web.icfes.gov.co/web/index.php?option=com_content&task=view&id=105&Itemid=128

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

López, A. y Janssen, G. (2010). Validation study of Colombia's ECAES English exam. *Lenguaje*, 38 (Nº 2), 423-448.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds.), *Test validity* (p.p. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associate.

_____ (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (p.p. 13-103). New York: ACE/ MacMillan.

Miles, M. B. y Huberman, A. M. (1994). Qualitative data analysis: an expanded sourcebook. Thousand Oaks, CA: Sage. (2nd ed).

Ministerio de Educación Nacional (2005). Colombia Bilingüe. *AlTablero* (37). Recuperado el 14 de Junio de 2008. Disponible en: http://www.mineducacion.gov.co/1621/article-97495.html

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Shepard, L. (1993). Evaluating test validity. *Review of research in education*, 19, 405-450.

Stratman, J. F. y Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: reflections on research methodology* (p.p. 89-111). Thousand Oaks, CA: Sage.

Tejada, H. y Castillo, N. (2010). El *Backwash Effect* o los efectos colaterales del examen ECAES, Prueba de inglés 2009. Un análisis crítico. *Lenguaje, 38* (Nº 2), 449-480

Anexo 1

OVERALL ALIGNMENT

Question 1: In your opinion, how well is the Saber 11 English Exam aligned with the CEFR?

Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned At All
0	0	0	0	0

Explain your response:

Question 2: In your opinion, how well is the Saber 11 English Exam aligned with the instruction at your institution?

Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned At All
0	0	0	0	0

Explain your response:

Question 3: In your opinion, what content is NOT assessed in the Saber 11 English exam that SHOULD BE assessed?

Anexo 2

Protocolo de Pensar en Voz Alta

Instrucciones:

En esta sesión te vamos a pedir que contestes la evaluación de uno de tus cursos pensando en voz alta. Esto quiere decir que a medida que vayas contestando la evaluación queremos que nos vayas verbalizando todo el proceso que usas para contestar cada una de las preguntas. Cada vez que contestes una pregunta en la evaluación, te voy a hacer unas preguntas complementarias para tener información sobre la claridad de las mismas. Cuando contestes todas las preguntas te voy a hacer una entrevista muy corta sobre tus impresiones sobre la evaluación y el proceso para completarla.

Si el participante se queda callado en alguna pregunta, usar cualquiera de las siguientes estrategias:

Preguntar

¿En qué te hace pensar esta pregunta?
¿En qué estás pensando en estos momentos?
¿Qué me puedes decir sobre esta pregunta?

Si el participante no dice mucho sobre alguna de las preguntas o no es muy claro, usar cualquiera de las siguientes estrategias:

Preguntar

¿Me puedes aclarar lo que acabas de decir?
¿Me puedes decir más sobre…?
¿Por qué dijiste que…?

Después de cada pregunta:

¿Qué te parece esta pregunta?
¿Crees que esta pregunta es clara?
¿Por qué sí? / ¿Por qué no?
¿Tienes algún otro comentario sobre esta pregunta?

Entrevista Retrospectiva

¿Qué te parece esta evaluación? ¿Cuál es tu impresión sobre esta evaluación?
¿Qué tan fácil o difícil fue contestar esta evaluación?
¿Qué fue lo que más te gustó de esta evaluación?
¿Qué fue lo que menos te gustó de esta evaluación?

Anexo 3

Encuesta – Estudiantes

El Centro de Evaluación del CIFE de la Universidad de los Andes está realizando un estudio de investigación que examina la validez del Examen de Estado Saber 11 de Inglés –ICFES– en los estudiantes de grado 11. Si quieres participar, por favor completa la siguiente encuesta. Esta encuesta es totalmente anónima y la información que des en ella se usará únicamente con fines académicos. A continuación encontrarás dos secciones y te tomará aproximadamente 20 minutos para completarlas.

1. Información personal

Edad: _____ Sexo: _____
Años en este colegio: _____

2. Información sobre el Examen de Inglés del ICFES

a. ¿Cómo te ayuda el colegio a prepararte para presentar el Examen de Inglés del ICFES?

_____
_____
_____
_____

b. ¿Qué haces por tu cuenta para prepararte para el Examen de Inglés del ICFES?

_____
_____
_____
_____

c. ¿Tienes alguna pregunta, comentario o inquietud sobre este estudio?

_____
_____
_____
_____

Muchas gracias por tu participación