

UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS
FACULTAD DE CIENCIAS SOCIALES
ESCUELA DE CIENCIAS PSICOLÓGICAS
MAESTRÍA EN PSICOMETRÍA Y EVALUACIÓN EDUCATIVA



TESIS

“Construcción de una prueba de inglés English as a Foreign Language (EFL) para valoración de competencia lectora en estudiantes de cuarto grado en escuelas públicas de la región #11 del Distrito Central de Francisco Morazán, Honduras para el año lectivo 2024”

Presentada por:

Suany Lisbeth Calderon Ruiz

Para optar al grado de Máster en Psicometría y Evaluación Educativa

Asesor Temático: Rigoberto de Jesús Padilla, MSc.

Asesor Metodológico: German Edgardo Moncada Godoy, PhD

Tegucigalpa, M.D.C., Honduras, C.A.

Abril, 2024

Tegucigalpa, M.D.C., 29 de Abril de 2025

Señores

Universidad Nacional Autónoma de Honduras

Coordinadores Maestría en Psicometría y Evaluación Educativa

Estimados señores:

Por este medio autorizo a la UNAH para que publique en su REVISTA DIGITAL, WEB, CD Y CUALQUIER OTRO MEDIO, y uso de beneficio sociales y académicos de la sociedad hondureña mi trabajo de tesis titulado: **“Construcción de una prueba de inglés English as a Foreign Language (EFL) para valoración de competencia lectora en estudiantes de cuarto grado en escuelas públicas de la región #11 del Distrito Central de Francisco Morazán, Honduras para el año lectivo 2024”**, el cual he presentado como parte de los requisitos para optar por la Maestría en Psicometría y Evaluación Educativa.

Atentamente,

Suany Lisbeth Calderon Ruiz

Cuenta No. PSE100208

AUTORIDADES UNIVERSITARIAS

Doctor Odir Fernando Flores
Rector

Doctora Lourdes Murcia
Vice – Rectora Académica

Abogado José Alexander Ávila
Secretaria General

Master Oscar Arquímedes Zelaya Villafranca
Director Sistema de Estudios de Postgrado

Master Carmen Julia Fajardo
Decana Facultad de Ciencias Sociales

Doctora María Victoria Ponce Mendoza
Coordinadora General Postgrados Facultad de Ciencias Sociales

Doctor German Edgardo Moncada Godoy
Coordinador Académico Maestría en Psicometría y Evaluación Educativa

Doctor Marcio Alexander Castillo Diaz
Coordinador de Investigación y Vinculación

Dedicatoria

Dedico este trabajo a Dios quien me ha permitido realizarlo concediéndome vida, salud, sabiduría, fortaleza, paciencia, los medios y las capacidades para cumplir a cabalidad este proyecto.

DEGT-UNAH

Agradecimiento

Agradezco a Dios mi Padre Celestial y a su Hijo Jesucristo, mi Señor y Salvador, en quien están escondidos todos los tesoros de la sabiduría y el conocimiento, que me ha concedido, en su infinita gracia, la porción de sabiduría necesaria para realizar esta tesis. Porque nada es imposible para Él, y al pedirle por sabiduría, concede abundantemente y sin reproche.

DEGT-UNAH

Resumen

El presente estudio tuvo como objetivo construir y analizar psicométricamente un instrumento para la evaluación de la competencia lectora en inglés como lengua extranjera (EFL) en estudiantes hondureños de cuarto grado. La validación del contenido se llevó a cabo mediante juicio de expertos, utilizando el coeficiente de validez de contenido (CVC), mientras que la validez de constructo se abordó a través de un análisis factorial exploratorio (AFE). Asimismo, se evaluó la consistencia interna mediante los coeficientes alfa de Cronbach y omega de McDonald. Finalmente, se aplicó la Teoría de Respuesta al Ítem (TRI), utilizando el modelo logístico de dos parámetros (2PL), con el fin de estimar los parámetros de discriminación y dificultad de cada ítem.

Los resultados indicaron una adecuada validez de contenido, aunque el AFE no evidenció una estructura factorial clara. Los coeficientes de confiabilidad fueron bajos, lo cual sugiere una limitada consistencia interna del instrumento en su versión actual. El análisis TRI permitió identificar ítems con buen comportamiento psicométrico, así como reactivos que requieren ajustes debido a parámetros atípicos. Se concluye que el instrumento constituye una base inicial sólida, aunque se requiere su depuración técnica para garantizar una medición válida, confiable y contextualizada de la competencia lectora en inglés.

Palabras clave: análisis psicométrico, competencia lectora, inglés como lengua extranjera, modelo 2PL, Teoría de Respuesta al Ítem.

Índices de Contenido

Dedicatoria.....	i
Agradecimiento	ii
Resumen	iii
Índice de Tablas y figuras	vi
Capítulo I.....	8
1. Planteamiento del Problema.....	8
1.1 Situación del problema.....	8
1.2 Objetivos de la Investigación	11
1.3 Justificación.....	12
CAPÍTULO II.....	14
2. Marco Teórico.....	14
2.1 El inglés como Lengua extranjera (EFL)	14
2.2 El inglés como segunda lengua (ESL)	15
2.3 Habilidades del ingles	16
2.4 La competencia Lectora en ingles.....	17
2.5 Modelo de Procesamiento de Lectura	17
2.6 Construcción de Pruebas.....	18
2.7 Tipos de pruebas	22
2.7.1 Pruebas estandarizadas.....	23
2.7.2 Pruebas de desempeño	31
2.7.3 Pruebas de conocimiento.....	33
2.8 Psicometría.....	35
2.8.1 Validez.....	36
2.8.2 Validez de Contenido	37
2.8.3 Metodología para validar contenidos	37
2.8.4 Validez de Contenido Mediante Juicio de Expertos	37
2.8.5.1 Supuestos del análisis factorial.....	42
2.8.5.2 Fases del Análisis Factorial.....	42
2.9 Fiabilidad.....	47
2.10 Teorías de Medición.....	50

2.10.1 Teoría Clásica de los Test	50
2.10.2 Teoría Respuesta al Ítem (TRI).....	51
2.10.2.1 Características desde la TRI.....	51
2.10.2.2 Supuestos desde la TRI	53
2.10.2.3 Modelos de la TRI.....	54
2.10.2.4 El modelo logístico de un parámetro	54
2.10.2.5 El modelo logístico de dos parámetros	55
Parámetro de Dificultad (b).....	56
2.10.2.6 El modelo logístico de tres parámetros	57
CAPÍTULO III	59
3. Marco Contextual.....	59
3.1 Evaluaciones Estandarizadas	59
3.2 Principales Pruebas para Evaluar la Lectura en inglés en Educación Primaria	60
3.3 Otras pruebas estandarizadas	61
3.4 Evaluaciones estandarizadas en Latinoamérica	63
3.2.5 Evaluaciones en la educación básica de Honduras	64
3.2.5.1 Evaluaciones Nacionales.....	64
3.2.5.2 Evaluaciones Internacionales	65
3.2.5.3 Evaluaciones en el Área de Inglés.....	65
CAPITULO IV	67
4. Metodología	67
4.1 Tipo de investigación	67
4.2 Población y Muestra.....	67
4.3 Variables e indicadores	68
4.4 Instrumento de medición.....	69
4.5 Plan de análisis de datos.....	70
4. Análisis de los Ítems mediante Teoría de Respuesta al Ítem (TRI).....	73
CAPÍTULO V.....	75
5. Resultados de la Investigación	75
5.1 Validez de contenido por jueces expertos	75
5.1.2 Descriptivos de los ítems del instrumento.....	76
5.1.3 Validez Estructural (Análisis Factorial Exploratorio (AFE)).....	77

5.1.4 Análisis de confiabilidad	78
5.1.5 Estimación de parámetros psicométricos desde la TRI y su función de información (2PL)	78
Tabla 3.....	¡Error! Marcador no definido.
CAPÍTULO VI	80
6. Discusión de los Resultados.....	81
6.1 Hallazgos Principales (validez por Jueces Expertos)	81
6.1.2 Validez Estructural (Análisis Factorial Exploratorio (AFE)).....	81
6.1.3 Confiabilidad (Consistencia interna).....	82
6.1.4 Discusión de los resultados del análisis TRI (modelo 2PL).....	83
6.2 Contraste de los principales hallazgos con la teoría.....	84
6.3 Implicaciones teóricas y prácticas de los resultados	85
Conclusiones	87
Recomendaciones	89
Referencias Bibliográficas	91
Declaración de Ética de Investigación.....	107
Anexos	108

Índice de Tablas

Tabla 1: Resumen del análisis de validez de contenido mediante juicio de expertos (CVC)	76
Tabla 2: Modelos factoriales exploratorios en función de cargas factoriales y presencia de cargas negativas	77
Tabla 3: Ítems con parámetros representativos y problemáticos del modelo 2PL, estimados para el instrumento de evaluación de competencia lectora en inglés (EFL).	79

Capítulo I

1. Planteamiento del Problema

El presente apartado expone los argumentos que justifican la necesidad de crear un instrumento, el cual ha sido diseñado para evaluar la competencia lectora adquirida a lo largo del primer ciclo escolar de la clase de inglés como lengua extranjera (EFL) en niños de 4to grado de educación primaria. Las ideas que se desarrollan son las siguientes: se describe la situación problemática que da origen al estudio, además se subraya que este instrumento es el primero en evaluar los aprendizajes adquiridos durante el primer ciclo escolar en competencia lectora del área de inglés como lengua extranjera y requirió validación por juicio de expertos, además de realizar análisis psicométricos, de tal manera que podrá garantizar el éxito académico en los estudiantes que se sometan a la misma en futuras aplicaciones.

1.1 Situación del problema

En Honduras, la evaluación del desempeño académico se ha centrado principalmente en áreas como español y matemáticas mediante pruebas estandarizadas implementadas desde 1997 (FEREMA, 2017). Sin embargo, no existe un instrumento validado que mida de manera específica la **competencia lectora en inglés como lengua extranjera (EFL)** en estudiantes de educación primaria. Esta ausencia limita la posibilidad de evaluar el nivel de comprensión lectora de los estudiantes en este idioma y, por ende, dificulta la toma de decisiones informadas para mejorar la enseñanza y el aprendizaje del inglés en el país (Brito et al. 2023).

A nivel internacional, existen diversas pruebas estandarizadas que evalúan habilidades lingüísticas en inglés como lengua extranjera (EFL), entre ellas el **Test of English as a Foreign Language (TOEFL Primary)**, el **Key English Test (KET)** y el **Preliminary English Test (PET)** de Cambridge. Estas pruebas incluyen componentes específicos de comprensión lectora en inglés y han sido diseñadas para distintos niveles escolares (Assessment, 2024; Wolf, 2024). En educación primaria, evaluaciones como **PIRLS (Progress in International Reading Literacy Study)** han medido la comprensión lectora, pero su aplicación en Honduras ha sido limitada y no considera el aprendizaje del inglés como lengua extranjera (Mullis et al. 2017). Un aspecto clave en la validez de estas pruebas

es la aplicación de métodos psicométricos rigurosos, tales como el Análisis Factorial Exploratorio (AFE) para validar la estructura de los ítems y la Teoría de Respuesta al Ítem (TRI) para analizar los parámetros de discriminación y dificultad de cada reactivo.

A pesar de los avances en la evaluación de EFL a nivel mundial, en el **contexto hondureño no existe una prueba psicométricamente validada** que mida la competencia lectora en inglés en la educación primaria. Hasta la fecha, la enseñanza del inglés se guía por los Estándares Educativos Nacionales de Inglés (EENI), los cuales establecen cuatro competencias fundamentales: Reading (lectura), Writing (escritura), Speaking (habla) y Listening (escucha) (Chamorro-Ortega et al., 2020). Sin embargo, no se ha desarrollado un instrumento específico que evalúe la competencia lectora en inglés basado en estándares psicométricos reconocidos, lo que dificulta la identificación del nivel real de los estudiantes y su preparación para el siguiente ciclo escolar.

Dada la creciente importancia del inglés como herramienta de comunicación global y su papel en el acceso a oportunidades académicas y laborales, diversos estudios han resaltado la necesidad de mejorar la enseñanza de este idioma en América Latina, implementando estándares educativos que favorezcan su aprendizaje (Sánchez-Jabba, 2013; Beltrán-Santoyo, 2021). En este sentido, la medición de la competencia lectora en inglés (EFL) requiere de instrumentos psicométricamente validados que permitan evaluar de manera confiable el desempeño de los estudiantes. Para que dicha prueba sea efectiva, es imprescindible que cumpla con los criterios psicométricos de validez y confiabilidad, garantizando así mediciones precisas y comparables (Muñiz, & Fonseca - Pedrero, 2019). Además de su aplicabilidad en el contexto hondureño, esta investigación busca generar evidencia empírica que permita fortalecer futuras iniciativas de evaluación en el ámbito de la enseñanza del inglés como lengua extranjera, facilitando la toma de decisiones pedagógicas informadas y la mejora de los programas educativos en la región (Mendoza Navas, 2024). Según el Instituto Nacional para la Evaluación de la Educación (INEE), (2019) la estandarización de instrumentos de evaluación localmente validados representa un avance significativo en la mejora del sistema educativo nacional, permitiendo una medición más precisa de las habilidades lectoras de los estudiantes y facilitando la implementación de estrategias de enseñanza alineadas con los resultados obtenidos en la prueba.

En este sentido, la presente investigación responde a la necesidad de desarrollar y validar un instrumento de medición que cumpla con criterios psicométricos rigurosos y que sea

culturalmente relevante para la población hondureña, permitiendo obtener inferencias válidas y confiables sobre el desempeño lector en inglés en niños de cuarto grado de educación primaria. Este estudio busca desarrollar una prueba de competencia lectora en inglés y someterla a un proceso de validación psicométrica que garantice su uso confiable en el contexto educativo hondureño (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

1.2 Objetivos de la Investigación

1.2.1 Objetivo General

Construir una prueba que permita evaluar la competencia lectora del inglés como lengua extranjera (EFL) en estudiantes de cuarto grado de educación básica primaria en escuelas públicas de Tegucigalpa.

1.2.2 Objetivos Específicos

- Analizar la validez de contenido para las pruebas del área de inglés (EFL) en estudiantes de primer ciclo de educación primaria a través de juicio de expertos.
- Calcular la validez estructural de la prueba de inglés (EFL) y determinar un diseño piloto por medio de análisis factorial exploratorio (AFE).
- Determinar la confiabilidad de los ítems y la estabilidad de la prueba del área de inglés (EFL) mediante el alfa de Cronbach y el Omega de McDonald.
- Determinar los parámetros psicométricos de discriminación y dificultad de la prueba del área de inglés /EFL) a través de la Teoría de Respuesta al ítem (TRI).

1.3 Justificación

En este espacio se describe el valor del conocimiento que aportará el estudio de construcción de la prueba de inglés como lengua extranjera (EFL) para valoración de competencia lectora en niños de educación primaria. Además de su utilidad en el fortalecimiento del sistema educativo de educación básica de Honduras asegurando la calidad de los estudiantes. Asimismo, aportando científicamente en la aplicación de procedimientos y metodologías de validación de instrumentos.

De acuerdo con (Alonso et al., 2021) en un mundo con tanta demanda laboral es pertinente hacer frente al aprendizaje de lenguas que sean y que tengan mayor relevancia tanto a nivel nacional como a nivel mundial.

Sobre esta idea se justifica la utilidad del estudio de construcción, ya que desde una perspectiva general, resulta fundamental determinar si un estudiante de primer ciclo de educación primaria **cumple con los estándares nacionales del área de inglés** establecidos por la **Secretaría de Educación de Honduras**, ya que estos criterios permiten evaluar el desarrollo de las competencias lingüísticas esperadas en esta etapa educativa (LPI-SEDUC-001-2015, 2015). Es por esto de la relevancia de la construcción de una prueba de evaluación de competencia lectora que surge con el propósito de mejorar y optimizar los procesos de enseñanza aprendizaje en la enseñanza del inglés como lengua extranjera.

El estudio de construcción del Test de Competencia lectora de Inglés como lengua extranjera (EFL) es relevante ya que la creación de una prueba de evaluación de competencia lectora en inglés como lengua extranjera, será un elemento que contribuirá para que los estudiantes que estén por realizar su paso al siguiente ciclo de educación primaria, considerando que el propósito no solo es que sirva como identificador, si no que contribuya a la construcción de conocimientos y con ello que se fundamente en pilares muy bien estructurados que sirvan como desarrollo de sus competencias (LPI-SEDUC-001-2015, 2015).

Este estudio aportará teórica y metodológicamente procedimientos claros para llevar a cabo estudios de validez por medio de un análisis psicométrico del Test de competencia lectora de inglés como lengua extranjera (EFL). Ya que la validez es el principio más importante de la psicometría (Martínez Arias et al., 2014). Como afirman (Cizek et al., 2010), “La validez es una de las mayores deidades en el panteón de la psicometría”.

Con este estudio también se pretende generar información que sea pertinente para las autoridades de la Secretaría de Educación la cual pueda hacer uso para mejora en el desarrollo de enseñanza – aprendizaje así contribuir en el progreso de cada estudiante o niño de acuerdo con las necesidades que permitan brindar una mejor educación y beneficiar en el avance de las competencias del educando el cual se tomara como proceso importante en la educación de calidad y así esta institución pueda contar con un instrumento que cumpla con propiedades psicométricas de validez y confiabilidad para nuestro contexto.

Finalmente argumentar que en el país **no se cuenta con una prueba** que permita evidenciar la confiabilidad y validez de la misma. Siendo esta una gran propuesta para evidenciar los aprendizajes obtenidos a lo largo del año escolar de los estudiantes del primer ciclo de educación primaria.

CAPÍTULO II

2. Marco Teórico

El siguiente apartado tiene como propósito mostrar algunas temáticas que respaldan la construcción de una prueba que mide la competencia lectora en inglés como lengua extranjera (EFL), se aborda el inglés como lengua extranjera (EFL) y como segunda lengua (ESL), la competencia lectora en inglés, las pruebas estandarizadas, validez de contenido, confiabilidad y las dos grandes teorías de medición: Teoría Clásica de los test (TCT) y Teoría de Respuesta al Ítem (TRI).

2.1 El inglés como Lengua extranjera (EFL)

Hoy en día, el idioma inglés es uno de los idiomas más solicitados y la demanda por saber este idioma cada vez es mayor. Las estadísticas mundiales de la Organización de la Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco Publishing, 2010) señalan al idioma inglés como la segunda lengua que más se practica. El inglés, español y francés están considerados como los lenguajes del mundo de los negocios; por tanto, para ingresar al mercado internacional en un futuro mediano, será necesario conocer y utilizar varios idiomas (Rueda Cataño & Wilburn Dieste, 2014). Al mismo tiempo aumenta las posibilidades de obtener un mejor empleo, oportunidades de viajar e intercambio de experiencias nuevas con personas y lugares. El proceso de aprendizaje del inglés como lengua extranjera generalmente se da dentro del aula de clase, lugar en el cual se realizan diferentes actividades de tipo controladas (Juan & García, 2012 citado por Beltran, 2017).

Diversos estudios han demostrado que el aprendizaje de una segunda lengua es más efectivo cuando ocurre antes de los 10 años, dado que la plasticidad cognitiva en esta etapa favorece la asimilación de estructuras gramaticales y fonológicas con mayor facilidad (Rueda Cataño & Wilburn Dieste, 2014). No obstante, la exposición al idioma en el sistema educativo público es limitada, oscilando entre 2 y 4 horas semanales, lo que restringe el desarrollo de competencias lingüísticas avanzadas (Beltrán-Santoyo, 2021). En contraste, los programas bilingües proporcionan entre 50 y 70% de instrucción en inglés, generando una mayor inmersión en el idioma y facilitando la adquisición de habilidades comunicativas según Chamorro-Ortega et al. (2020). Adicionalmente, el contexto sociocultural juega un papel

determinante en la consolidación del aprendizaje del inglés; los niños con mayor acceso a materiales digitales, televisión, y plataformas interactivas en inglés presentan un desarrollo más acelerado en comparación con aquellos cuya exposición se limita al aula (Sánchez-Jabba, 2013). Investigaciones sobre adquisición de segundas lenguas indican que los niños que inician el aprendizaje del inglés entre los 5 y 8 años tienen mayores probabilidades de alcanzar una competencia lectora y oral funcional en edades posteriores según Chamorro-Ortega et al. (2020). Sin embargo, el nivel de dominio alcanzado depende del modelo de enseñanza y la frecuencia de exposición; en el sistema público, la consolidación del inglés generalmente ocurre en la educación secundaria, mientras que, en instituciones bilingües, los estudiantes pueden desarrollar fluidez comunicativa desde la educación primaria de acuerdo con Beltrán-Santoyo et al. (2021).

2.2 El inglés como segunda lengua (ESL)

La etapa idónea para aprender una segunda lengua es la infancia, dada la plasticidad del cerebro y la falta de especialización cortical que caracteriza a esta etapa (Rueda Cataño & Wilburn Dieste, 2014). Las nuevas teorías lingüísticas para la enseñanza del idioma inglés a temprana edad, están retomando la conciencia fonológica (phonological awareness) como una herramienta que abona, primero, a los conocimientos de los docentes, y luego al proceso enseñanza-aprendizaje de la lecto-escritura en la adquisición de otra lengua; la conciencia fonológica se define como la habilidad que posibilita a los niños a reconocer, identificar, deslindar, manipular deliberadamente y actuar con los sonidos o fonemas que componen las palabras. Los niños no la adquieren en forma espontánea; ésta debe aprenderse y su desarrollo resulta imprescindible antes de que inicie la enseñanza gráfica del alfabeto.

Independientemente de la variedad de idiomas que se están volviendo una forma de comunicación más común, el inglés ocupa su lugar como una de las formas de comunicación predominantes en el mundo y su influencia se extienden hasta más de 2000 millones de personas en todo el mundo según Beltrán-Santoyo et al. (2021). Por otro lado, Chamorro-Ortega et al. (2020) establece que para aprender un idioma es necesario desarrollar las cuatro habilidades lingüísticas sin dejar de lado ninguna de estas y que cuando una persona empieza a aprender un idioma desde su infancia lo adquirirá como segunda lengua.

Hoy día, la originalidad y modernidad de los métodos para la enseñanza de una segunda lengua está en función de la práctica oral y comunicativa del idioma, y el ejercicio de la gramática; por ello resulta trascendental el estudio de las normas y principios que regulan la estructura y uso de un lenguaje, así como la gramática, que conlleva a la mejora en la comunicación y a un conocimiento más profundo del segundo idioma (Rueda Cataño & Wilburn Dieste, 2014).

La enseñanza de un segundo idioma, que inicia desde el tercer grado de preescolar (preparatoria) y constituye un verdadero y complejo desafío, ya que ni la alfabetización de la lengua materna, ni el aprendizaje de un idioma no nativo, son procesos espontáneos, es decir, su adquisición requiere de una intervención pedagógica. Este nivel inicial busca introducir a los estudiantes en el idioma mediante actividades básicas que fomenten el reconocimiento auditivo, la pronunciación de palabras simples y el desarrollo de vocabulario elemental. Esta base inicial permite una transición más sólida hacia los niveles superiores de educación primaria, donde se espera que los estudiantes fortalezcan habilidades lingüísticas más complejas, como la comprensión lectora y la escritura básica en inglés.

Aunque existen diferencias en la implementación según el contexto escolar, este inicio temprano es clave, pues aprovecha la plasticidad del cerebro infantil y su capacidad para adquirir un segundo idioma de manera más eficiente en comparación con etapas posteriores (Rueda Cataño & Wilburn Dieste, 2014). En este sentido identificamos un continuo educativo, que debe ir de la mano hasta el término de la secundaria, en el que se procure una secuencia y progresión entre los diversos niveles y modalidades de educación que se ofrecen, de manera que dé continuidad al desarrollo de los principios, valores y procesos que queremos promover (Rueda Cataño & Wilburn Dieste, 2014). De acuerdo con la Secretaría de Educación de Honduras (LPI-SEDUC-001-2015, 2015), la enseñanza del inglés en el sistema público comienza en el tercer grado de preescolar (Prebásica III, 5 años), mientras que en las instituciones privadas y bilingües la instrucción inicia incluso en Prebásica I (3 años) con mayor intensidad y frecuencia.

2.3 Habilidades del inglés

Chamorro et al. (2020) argumenta que para aprender un idioma correctamente debemos desarrollar las cuatro habilidades básicas: habilidad de hablar, escuchar, escribir y leer. Es

decir, se debe dominar la expresión oral, comprensión auditiva, expresión escrita y comprensión lectora, siendo esta última la variable principal a analizar en esta investigación.

2.4 La competencia Lectora en inglés

La competencia lectora en inglés se define como la capacidad del estudiante para identificar palabras, comprender el significado de frases y extraer información de textos simples en el idioma de acuerdo con Grabe y Stoller (2002).

Según McNamara (2004) la competencia lectora en inglés se ha convertido en un pilar fundamental dentro del aprendizaje de una segunda lengua. Diversas investigaciones han demostrado que el desarrollo de habilidades de lectura en inglés es un proceso complejo, influenciado por factores cognitivos, lingüísticos y motivacionales. Solé (1998) establece que la comprensión lectora es un proceso dinámico que involucra la interacción entre el lector y el texto. Por otro lado, Guarín & Ramírez (2017) afirman que este proceso depende de diversos factores, como el conocimiento previo del lector, sus objetivos al leer y la motivación para comprender la información. En el primer ciclo de primaria, esta habilidad se desarrolla progresivamente mediante el reconocimiento de palabras familiares, la lectura de oraciones cortas y la asociación de imágenes con significado textual (Ehri, 2005).

De acuerdo con Kintsch y van Dijk (1983), la comprensión lectora se desarrolla en dos niveles:

- **Comprensión literal:** Identificación de información explícita en el texto.
- **Comprensión inferencial:** Capacidad para deducir significados a partir del contexto.

Mientras que (Ramírez, 2017; Llorens, 2015) establecen que, en el caso del aprendizaje de inglés como lengua extranjera, se identifican distintos niveles de comprensión lectora: literal, inferencial y crítica.

2.5 Modelo de Procesamiento de Lectura

El modelo de **procesamiento de lectura** de Grabe & Stoller (2002) establece que los niños deben desarrollar varias habilidades para lograr una lectura efectiva en inglés:

- **Decodificación:** Identificación de palabras y estructuras gramaticales básicas.
- **Fluidez:** Capacidad de leer con ritmo y precisión.

- **Comprensión literal:** Identificación de ideas principales en textos simples.
- **Inferencia:** Uso del contexto y apoyo visual para deducir significados.

La evaluación de la competencia lectora en inglés requiere el diseño de pruebas que midan de manera válida y confiable las habilidades de comprensión de los estudiantes (Alderson, 2000). Para ello, es fundamental aplicar principios de construcción de pruebas que garanticen la calidad psicométrica del instrumento, lo que se abordará en el siguiente apartado.

2.6 Construcción de Pruebas

La construcción de pruebas es un proceso complejo que requiere una planificación meticulosa para garantizar su validez y confiabilidad. Según Anastasi y Urbina (1998), el desarrollo de una prueba implica la identificación del constructo, la selección de ítems representativos y la aplicación de análisis estadísticos para evaluar su calidad psicométrica. De acuerdo con Muñiz (2018), la validez de una prueba no es un atributo estático, sino un proceso de acumulación de evidencia basado en diferentes fuentes, como el análisis factorial exploratorio y confirmatorio, así como en la Teoría de Respuesta al Ítem (TRI), que permite examinar el funcionamiento de cada ítem y su contribución a la medición del constructo. Por su parte, Downing y Haladyna (2006) sostienen que la construcción de pruebas debe seguir principios de diseño estructurado, asegurando la representatividad del contenido y la equidad en la evaluación. Asimismo, la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME) (2014) enfatizan la necesidad de establecer estándares éticos y técnicos para el desarrollo de pruebas, asegurando su imparcialidad, validez y utilidad en contextos educativos y de investigación. En este sentido, Prieto y Delgado (2010) subrayan la importancia de la calidad de los ítems en función de su discriminación y dificultad, destacando que la precisión en la medición depende de un adecuado diseño y análisis estadístico de los instrumentos. Así, el proceso de construcción de pruebas no solo implica criterios estadísticos rigurosos, sino también consideraciones teóricas, éticas y prácticas que garanticen su eficacia en la evaluación de los constructos medidos.

Para ello la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education

(NCME) (2018) establecen estándares que guían el diseño y la validación de pruebas educativas y psicológicas, es necesario seguir un modelo metodológico que estructure cada una de las fases del desarrollo del instrumento, permitiendo su validación.

En esta etapa, es fundamental considerar los siguientes aspectos:

1. Propósito y Uso de la Prueba

Toda prueba debe desarrollarse con un objetivo claro. Se pueden clasificar en:

- **Pruebas diagnósticas:** Identifican fortalezas y debilidades en un área específica.
- **Pruebas formativas:** Monitorean el aprendizaje en proceso y proporcionan retroalimentación.
- **Pruebas sumativas:** Evalúan el rendimiento global después de un período de instrucción.
- **Pruebas normativas:** Comparan el desempeño de un examinado con un grupo de referencia.
- **Pruebas criterioles:** Determinan si un individuo ha alcanzado un nivel específico de competencia.

La claridad en el propósito de la prueba influye en la selección de ítems y en la interpretación de los puntajes.

2. Definición del Constructo y Tabla de Especificaciones

El **constructo** es la habilidad, conocimiento o rasgo que la prueba pretende medir (APA, AERA, NCME, 2014), su correcta definición permite delimitar los aspectos relevantes y evitar la inclusión de contenido irrelevante. Para ello, se recomienda elaborar una tabla de especificaciones. Según Muñiz (2018) la tabla de especificaciones es un instrumento esencial en la construcción de pruebas estandarizadas, ya que permite estructurar la distribución de los ítems en función de las dimensiones del constructo y los contenidos educativos relevantes en la cual detalla:

- Dimensiones o subcomponentes del constructo.
- Cantidad de ítems por cada dimensión.
- Distribución de los niveles de complejidad (por ejemplo, basado en la taxonomía de Bloom).

Esta planificación ayuda a asegurar la representatividad del contenido en la prueba y la alineación con los objetivos de evaluación.

3. Desarrollo de Ítems

El desarrollo de ítems es una fase crítica en la construcción de pruebas, ya que la calidad de los reactivos impacta directamente en la validez y confiabilidad de la evaluación.

Tipos de Ítems

Los ítems pueden diseñarse en distintos formatos según el objetivo de evaluación:

- **Opción múltiple:** Se compone de un enunciado y varias alternativas de respuesta, de las cuales solo una es correcta. Se recomienda utilizar distractores plausibles para evitar respuestas por eliminación.
- **Respuesta corta:** Requiere que el examinado escriba una breve respuesta. Útil para evaluar conocimientos puntuales.
- **Desarrollo o ensayo:** Exige que el examinado elabore una respuesta extensa, evaluando argumentación y pensamiento crítico.
- **Escalas de Likert:** Miden actitudes o percepciones a través de opciones de respuesta graduadas (por ejemplo, “Totalmente en desacuerdo” a “Totalmente de acuerdo”).
- **Reactivos basados en desempeño:** Incluyen tareas o problemas prácticos que requieren aplicación del conocimiento en contextos reales.

4. Revisión y Refinamiento de Ítems

Los ítems deben ser revisados antes de ser aplicados en una prueba piloto. Se recomienda:

- **Juicio de expertos:** Un grupo de especialistas analiza la claridad, relevancia y adecuación de los ítems.
- **Análisis de contenido:** Se verifica que los ítems representen adecuadamente el constructo definido.
- **Prueba piloto:** Se aplica la prueba en un grupo reducido para analizar el comportamiento de los ítems mediante estadísticas psicométricas.

Por otro lado, Muñiz y Fonseca-Pedrero (2019) en su artículo **“Diez pasos para la construcción de un test”** presentan un marco sistemático para la construcción de pruebas psicométricas con base en estándares internacionales. Este modelo es aplicable tanto en la evaluación de constructos psicológicos como en pruebas educativas, asegurando la calidad de los instrumentos de medición.

1. Definición del Marco General

El primer paso en la construcción de un test consiste en justificar su necesidad, definir su objetivo y especificar su contexto de aplicación. Es fundamental delimitar la variable a medir, la población objetivo y el tipo de aplicación (individual o colectiva). Además, se debe establecer el formato de administración (papel, digital, etc.) y las decisiones que se tomarán a partir de los resultados obtenidos (Muñiz, & Fonseca - Pedrero, 2019).

2. Definición de la Variable a Medir

En esta fase, es crucial realizar una definición operativa de la variable evaluada para garantizar su medición empírica. Se debe identificar con precisión qué aspectos forman parte del constructo y cuáles no, evitando tanto la infrarrepresentación como la sobrerrepresentación de la variable. La consulta a expertos y la revisión de la literatura ayudan a definir con claridad los indicadores más representativos de la variable en cuestión.

3. Especificaciones del Test

Aquí se establecen los detalles técnicos de la prueba, incluyendo el número de ítems, su formato, el tiempo de aplicación y los requerimientos de accesibilidad. Se deben considerar aspectos como el tipo de respuestas permitidas, la administración en poblaciones con necesidades especiales y la forma en que se controlarán los sesgos de respuesta (Muñiz, & Fonseca - Pedrero, 2019).

4. Construcción de los Ítems

Los ítems son la base de cualquier prueba y deben diseñarse cuidadosamente para garantizar su validez. Se recomienda seguir principios como la claridad, representatividad y pertinencia en la redacción de los reactivos. Además, si los ítems provienen de pruebas existentes en otro idioma o contexto, deben adaptarse siguiendo directrices internacionales de traducción y equivalencia cultural (Hambleton R. K., 2005).

5. Edición del Test

La primera versión del test debe ser cuidadosamente revisada antes de su aplicación piloto. En esta fase, se optimiza la presentación del material, asegurando que los ítems sean comprensibles y estén organizados de manera lógica. Un diseño claro y profesional del test mejora la experiencia del examinado y reduce posibles errores de aplicación.

6. Estudios Piloto

La prueba debe aplicarse a un grupo reducido de personas con características similares a la población objetivo para detectar posibles errores. Se pueden realizar estudios piloto

cualitativos (grupos de discusión sobre la claridad de los ítems) y cuantitativos (análisis psicométricos preliminares). Los resultados del piloto permiten ajustar los ítems y mejorar la calidad del test antes de su aplicación definitiva (Muñiz, & Fonseca - Pedrero, 2019).

7. Selección de Otros Instrumentos de Medida

Para validar el test, es recomendable utilizar otras pruebas previamente validadas que midan constructos similares o relacionados. Esto permite establecer evidencias de validez convergente y discriminante, fundamentales para garantizar la precisión de las interpretaciones derivadas de la prueba.

8. Aplicación de la Prueba

En esta etapa, se administra la prueba a una muestra representativa de la población objetivo. Es fundamental seguir un protocolo estandarizado para garantizar condiciones homogéneas en la administración del test y minimizar errores de medición. Además, se debe garantizar la calidad de la base de datos generada (Muñiz, & Fonseca - Pedrero, 2019).

9. Análisis de Propiedades Psicométricas

Una vez aplicada la prueba, se evalúan sus propiedades psicométricas mediante técnicas estadísticas. Se analizan la validez de contenido, la estructura factorial, la confiabilidad (Alfa de Cronbach, Omega de McDonald), y el funcionamiento diferencial de los ítems (DIF) para garantizar que el test mida adecuadamente el constructo de interés sin sesgos.

10. Desarrollo de la Versión Final

Tras el análisis de los resultados, se ajusta la versión definitiva del test, documentando en un manual técnico sus propiedades psicométricas, instrucciones de aplicación y pautas para la interpretación de puntajes. Los autores destacan que la validación de una prueba es un proceso continuo, ya que nuevos estudios pueden aportar más evidencia sobre su validez y aplicabilidad en diferentes contextos (Muñiz, & Fonseca - Pedrero, 2019).

2.7 Tipos de pruebas

El proceso de construcción de pruebas es un componente fundamental en la evaluación educativa y profesional, ya que garantiza la validez, confiabilidad y utilidad de los instrumentos diseñados para medir el aprendizaje y las competencias de los individuos (Anastasi & Urbina, 1998; Muñiz, 2018). Sin embargo, la naturaleza y el propósito de las

pruebas varían según el tipo de evaluación que se desee realizar. En este sentido, las pruebas pueden clasificarse en **pruebas de conocimiento**, que miden la adquisición de información y comprensión conceptual; **pruebas de desempeño**, que evalúan la aplicación práctica de habilidades en contextos auténticos; y **pruebas estandarizadas**, diseñadas para comparar el rendimiento de los evaluados con una norma previamente establecida (Brookhart, 2013; Darling-Hammond & Adamson, 2014). Cada una de estas pruebas responde a objetivos específicos y requiere un diseño metodológico particular que permita garantizar mediciones precisas y relevantes para la toma de decisiones en los ámbitos educativo, profesional y social (Pellegrino et al. 2011).

2.7.1 Pruebas estandarizadas

Las pruebas estandarizadas internacionales constituyen herramientas fundamentales para la evaluación comparativa del desempeño académico en distintos países. Desarrolladas por agencias y organizaciones internacionales, estos instrumentos de medición estandarizados permiten recopilar información sobre los aprendizajes conceptuales y procedimentales en diversas áreas del conocimiento, además de analizar los factores contextuales que influyen en el aprendizaje de los estudiantes de acuerdo con Sánchez (2020). Estas evaluaciones no solo identifican habilidades, destrezas y áreas de mejora en los estudiantes, sino que también ofrecen información clave para fortalecer el conocimiento y orientar estrategias que favorezcan un mejor rendimiento académico. En este sentido, las pruebas de gran escala, tanto nacionales como internacionales, deben contar con un diseño sólido que garantice la validez de los datos obtenidos, permitiendo así aportar información relevante para el beneficio del sistema educativo y de sus estudiantes.

Asimismo, este tipo de instrumentos de evaluación presentan una notable versatilidad, ya que pueden aplicarse a una muestra representativa de la población o a la totalidad de esta, convirtiéndose en un censo. Esto posibilita el desarrollo de análisis estadísticos sobre factores contextuales, como las condiciones sociales y económicas de los participantes, lo que enriquece la interpretación de los resultados obtenidos. Además, estas evaluaciones pueden incluir ítems diseñados para medir aspectos cognitivos, tales como habilidades, competencias y actitudes, proporcionando así una visión más integral del nivel educativo y la calidad del sistema de enseñanza en cada contexto afirma Gonzales Acosta (2019). Una

de las principales ventajas de las pruebas estandarizadas radica en su capacidad para ofrecer datos objetivos que permitan la identificación de áreas de mejora y la implementación de estrategias dirigidas al fortalecimiento del aprendizaje en las distintas áreas evaluadas.

De acuerdo con Popham (2017), estas pruebas permiten obtener mediciones objetivas del aprendizaje estudiantil y facilitan la comparación entre distintos sistemas educativos. Desarrolladas por agencias y organizaciones internacionales, estos instrumentos de medición estandarizados recopilan información sobre los aprendizajes conceptuales y procedimentales en diversas áreas del conocimiento, además de analizar los factores contextuales que influyen en el desempeño de los estudiantes según Sánchez (2020). Tal como lo afirma McMillan (2015), una de sus principales ventajas es la posibilidad de identificar fortalezas y debilidades en la formación de los estudiantes, lo que contribuye a la toma de decisiones informadas para mejorar la calidad educativa. En este sentido, las pruebas de gran escala, tanto nacionales como internacionales, deben contar con un diseño psicométrico riguroso que garantice la validez y confiabilidad de los datos obtenidos (APA, AERA, NCME, 2014), asegurando que los resultados reflejen con precisión el nivel de desempeño de los evaluados. Además, estos instrumentos de evaluación presentan una notable versatilidad, ya que pueden aplicarse a una muestra representativa de la población o a la totalidad de esta, convirtiéndose en un censo asegura Gonzales Acosta (2019). Esto permite el desarrollo de análisis estadísticos avanzados sobre factores contextuales, como las condiciones sociales y económicas de los participantes, lo que enriquece la interpretación de los resultados y su relación con las inequidades educativas (OECD, 2019). Asimismo, como señalan Brookhart y Nitko (2019), las pruebas estandarizadas pueden incluir ítems diseñados para evaluar aspectos cognitivos, tales como habilidades, competencias y actitudes, proporcionando una visión más integral del nivel educativo y la calidad del sistema de enseñanza en cada contexto. No obstante, es fundamental considerar las limitaciones de estas pruebas, como su dependencia de formatos de respuesta cerrados, lo que puede restringir la evaluación de habilidades complejas según Stiggins (2008). Aun así, una de sus principales ventajas radica en su capacidad para ofrecer datos comparables y objetivos que permitan la identificación de áreas de mejora, el establecimiento de políticas educativas y la implementación de estrategias dirigidas al fortalecimiento del aprendizaje en las distintas áreas evaluadas.

Existen diversas pruebas estandarizadas a nivel internacional, nacional y regional, diseñadas para evaluar distintas áreas del conocimiento y habilidades en diferentes contextos. Estas

pruebas han sido sometidas a rigurosos análisis psicométricos para garantizar su validez y confiabilidad en la medición de las competencias evaluadas. A continuación, se presentan algunas de las más relevantes:

- El **Young Learners English (YLE)** es una prueba estandarizada desarrollada por Cambridge Assessment English con el propósito de evaluar la competencia en inglés de niños en educación primaria. Su diseño responde a la necesidad de contar con instrumentos de evaluación que permitan medir de manera válida y confiable las habilidades lingüísticas en edades tempranas, asegurando que el contenido evaluado sea accesible y adecuado para el nivel de desarrollo cognitivo de los estudiantes (Cambridge Assessment English, 2021). La prueba se estructura en tres niveles progresivos: Pre A1 Starters, dirigido a estudiantes en etapa inicial de aprendizaje del idioma; A1 Movers, que evalúa habilidades básicas de comprensión y producción oral y escrita; y A2 Flyers, que mide la capacidad del estudiante para comprender y expresarse en contextos más complejos (Cambridge Assessment English, 2021). Cada nivel evalúa las cuatro macro habilidades lingüísticas fundamentales: comprensión auditiva, lectura, escritura y expresión oral, mediante actividades diseñadas para simular el uso real del idioma en situaciones comunicativas propias del entorno infantil (Taylor, 2013).

Desde una perspectiva psicométrica, el YLE es sometido a rigurosos procedimientos de validación para garantizar su fiabilidad y validez como instrumento de evaluación. En términos de validez de contenido, la prueba está alineada con los descriptores del Marco Común Europeo de Referencia para las Lenguas (MCER), lo que permite asegurar que los ítems reflejan de manera precisa las habilidades lingüísticas establecidas para cada nivel de competencia (Weir, 2005). Asimismo, se emplea el coeficiente alfa de Cronbach para evaluar la consistencia interna de los ítems dentro de cada sección del examen, garantizando que los reactivos midan de manera homogénea los constructos teóricos subyacentes (Kunnan, 2018). Además, la prueba incorpora modelos psicométricos basados en la Teoría de Respuesta al Ítem (TRI), lo que permite analizar la dificultad y la capacidad discriminativa de cada ítem, optimizando la precisión en la estimación del nivel de competencia de los examinados (Alderson, J. C., 2018). Finalmente, previo a su implementación, el YLE es sometido a estudios piloto en poblaciones representativas,

asegurando la adecuada calibración de los ítems y su equidad en la evaluación de los estudiantes (English., Cambridge Assessment, 2021).

- **PISA (Programme for International Student Assessment):** El Programa para la Evaluación Internacional de los Estudiantes (**PISA**) es desarrollado por **la Organización para la Cooperación y el Desarrollo Económicos (OCDE)** y mide competencias en lectura, matemáticas y ciencias en estudiantes de 15 años (OECD, 2019). Su objetivo es evaluar el nivel de desempeño de los estudiantes en la resolución de problemas y la aplicación de conocimientos en contextos reales. PISA es una prueba estandarizada a nivel internacional, con altos niveles de validez y confiabilidad. Se basa en modelos de **Teoría de Respuesta al Ítem (TRI)** para garantizar la comparabilidad de los resultados entre diferentes países y poblaciones (OECD, 2019). Aunque no es una prueba específica de inglés, la sección de lectura evalúa la comprensión de textos en contextos académicos, lo que puede relacionarse con la adquisición de habilidades lingüísticas en un entorno de EFL.
- **PIRLS (Progress in International Reading Literacy Study):** El Estudio Internacional de Progreso en Comprensión Lectora (**PIRLS**) evalúa la capacidad de comprensión lectora en estudiantes de 4.º grado a nivel internacional (Mullis et al., 2017). PIRLS se basa en modelos de validez de contenido y de constructo, asegurando que mide adecuadamente la comprensión lectora en distintos idiomas. La confiabilidad de la prueba es alta debido a su enfoque en la comparación de tendencias a lo largo del tiempo (Mullis et al., 2017). PIRLS es particularmente útil en contextos de enseñanza del inglés como lengua extranjera, ya que permite evaluar el desarrollo de la comprensión lectora, una habilidad clave en el aprendizaje de **EFL**.
- **ICILS (International Computer and Information Literacy Study):** El Estudio Internacional de Alfabetización Digital e Informática (**ICILS**) mide competencias digitales y alfabetización informática en estudiantes de **8.º grado**, con un enfoque en la capacidad de manejar información y comunicarse en entornos digitales (Fraillon et al., 2019). ICILS se fundamenta en modelos de medición psicométrica que garantizan la validez y confiabilidad de los resultados. Se aplican análisis de dificultad y discriminación para asegurar una evaluación precisa (Fraillon et al., 2019). En un mundo digitalizado, el aprendizaje del inglés está cada vez más mediado por herramientas digitales. ICILS proporciona información sobre

cómo los estudiantes interactúan con textos en inglés en plataformas digitales, lo que es clave para su desarrollo en **EFL**.

- **TOEFL (Test of English as a Foreign Language):** El **Test of English as a Foreign Language (TOEFL iBT)** es una evaluación estandarizada desarrollada por el **Educational Testing Service (ETS)** con el propósito de medir la competencia en inglés de hablantes no nativos en contextos académicos (ETS., 2021). Su diseño y validación responden a estrictos criterios psicométricos, asegurando que la prueba proporcione mediciones válidas, confiables y comparables entre distintos grupos de examinados. Para ello, el TOEFL iBT emplea la **Teoría de Respuesta al Ítem (TRI)** como modelo de análisis, permitiendo la calibración de los reactivos con base en tres parámetros fundamentales: **dificultad (b)**, que determina el nivel de habilidad necesario para responder correctamente; **discriminación (a)**, que evalúa la capacidad de un ítem para diferenciar entre examinados con distintos niveles de competencia lingüística; y **pseudoazar o adivinanza (c)**, que controla la probabilidad de respuestas correctas obtenidas al azar (ETS., 2021). Adicionalmente, el examen ha sido sometido a **análisis factorial confirmatorio (AFC)** para verificar la estructura subyacente de la prueba y garantizar que las dimensiones evaluadas reflejen constructos teóricamente sólidos de adquisición del lenguaje (ETS., 2021). Asimismo, la confiabilidad de las puntuaciones es evaluada mediante el **coeficiente alfa de Cronbach**, asegurando una adecuada consistencia interna en las diferentes secciones del examen. Estudios de **validez predictiva** han demostrado una correlación significativa entre las puntuaciones del TOEFL iBT y el desempeño académico de los examinados en instituciones de educación superior, lo que refuerza su utilidad como predictor del éxito académico en entornos angloparlantes (ETS., 2021). En el ámbito de la evaluación de **Asistentes de Enseñanza Internacionales (ITAs)** en universidades de Estados Unidos, se han establecido **puntuaciones de corte** basadas en modelos psicométricos que permiten categorizar a los candidatos según su suficiencia lingüística para desempeñar funciones docentes en contextos universitarios (ETS., 2021). La implementación de estos procesos metodológicos respalda la validez del TOEFL iBT como instrumento de medición estandarizado y refuerza su papel en la evaluación de competencia lingüística en el ámbito académico.
- **SAT (Scholastic Assessment Test) y ACT (American College Testing):** El **SAT** y el **ACT** son exámenes utilizados en **Estados Unidos** para evaluar la preparación de los estudiantes

que ingresan a la educación superior. Ambas pruebas miden habilidades en lectura, escritura, matemáticas y razonamiento científico (College Board, 2020). Tanto el SAT como el ACT han pasado por extensos procesos de validación para garantizar su confiabilidad y equidad. Se basan en análisis psicométricos rigurosos para ajustar la dificultad de los ítems (College Board, 2020). Aunque no están diseñadas específicamente para evaluar inglés como segunda lengua, incluyen secciones de comprensión lectora y escritura que pueden ser un reto para hablantes no nativos. Son utilizadas por instituciones educativas para medir el nivel de inglés académico de estudiantes internacionales.

- El **Test for the Assessment of the English Language (TAEI)** es una **prueba estandarizada** diseñada para evaluar la competencia comunicativa en inglés en contextos profesionales y educativos (Bazaga, F., & Marín, F., 2003). Esta evaluación mide cuatro habilidades lingüísticas fundamentales: **comprensión lectora léxico-gramática, expresión escrita, expresión oral y comprensión auditiva**, proporcionando una puntuación global y por destrezas (Bazaga, F., & Marín, F., 2003). La construcción del TAEI siguió rigurosos procesos psicométricos, incluyendo estudios de **validez de constructo**, mediante el cual se confirmó que las escalas evaluadas corresponden a las dimensiones teóricas propuestas en el modelo de evaluación del idioma (Bazaga, F., & Marín, F., 2003). Además, se realizaron análisis de **validez de criterio**, comparando los resultados del TAEI con otras pruebas reconocidas para garantizar su utilidad predictiva (Bazaga, F., & Marín, F., 2003). La fiabilidad del test se estableció a través del **coeficiente alfa de Cronbach**, asegurando una alta consistencia interna entre los ítems de cada escala, y mediante **análisis factorial exploratorio (AFE)** y **confirmatorio (AFC)**, los cuales validaron la estructura de la prueba y la relación entre los ítems y los constructos teóricos medidos (Bazaga, F., & Marín, F., 2003). Finalmente, el test ha sido **bareado con una muestra representativa**, permitiendo la interpretación de sus resultados en función de normas poblacionales y su correspondencia con los niveles del **Marco Común Europeo de Referencia para las Lenguas (MCER)**, facilitando así su uso en procesos de selección de personal, evaluación académica y orientación lingüística (Bazaga, F., & Marín, F., 2003).
- El **Test for the Assessment of the English Language in the School (TAEIIS)** es una **prueba estandarizada** diseñada para evaluar la competencia en inglés de estudiantes de entre **11 y 16 años**, abarcando niveles educativos desde **6º de Primaria hasta 2º de**

Bachillerato (Bazaga, F., & Marín, F., 2003). Este instrumento evalúa cuatro dimensiones del aprendizaje del idioma: **comprensión lectora, expresión escrita, comprensión auditiva y expresión oral**, proporcionando resultados cuantificables y normativos que permiten ubicar a los evaluados en distintos niveles de competencia lingüística (Bazaga, F., & Marín, F., 2003). Para garantizar la validez y confiabilidad de la prueba, el **Taelis** ha sido sometido a rigurosos análisis psicométricos. Su estructura se ha validado mediante **Análisis Factorial Exploratorio (AFE)** y **Análisis Factorial Confirmatorio (AFC)**, permitiendo verificar que los ítems incluidos reflejen constructos teóricos sólidos de evaluación del idioma (Bazaga, F., & Marín, F., 2003). Asimismo, la prueba ha sido calibrada utilizando la **Teoría de Respuesta al Ítem (TRI)**, la cual permite analizar parámetros fundamentales como la **dificultad (b)**, que determina el nivel de habilidad requerido para responder correctamente; la **discriminación (a)**, que mide la capacidad del ítem para diferenciar entre distintos niveles de competencia lingüística; y el **pseudoazar o adivinanza (c)**, que ajusta la probabilidad de respuestas correctas obtenidas al azar (Bazaga, F., & Marín, F., 2003). Además, la confiabilidad del Taelis se ha establecido a través del **coeficiente alfa de Cronbach**, garantizando la consistencia interna de las distintas secciones de la prueba (Bazaga, F., & Marín, F., 2003). Finalmente, la evaluación ha sido **baremada con una muestra representativa**, permitiendo interpretar los resultados en relación con normas poblacionales y alinearlos con el **Marco Común Europeo de Referencia para las Lenguas (MCER)**, lo que refuerza su utilidad en la evaluación de la competencia lingüística en inglés en el ámbito escolar (Bazaga, F., & Marín, F., 2003).

- El **Duolingo English Test (DET)** es una evaluación estandarizada en línea diseñada para medir la competencia en inglés de hablantes no nativos. Su principal ventaja es la accesibilidad, ya que puede realizarse desde cualquier lugar con conexión a internet y sin necesidad de acudir a un centro de evaluación. El DET es un examen adaptativo, lo que significa que la dificultad de las preguntas varía en función de las respuestas del examinado, permitiendo una evaluación más eficiente del nivel de competencia lingüística del candidato (Duolingo, 2023). El examen tiene una duración aproximada de 60 minutos y evalúa las cuatro habilidades lingüísticas esenciales: comprensión lectora, comprensión auditiva, expresión oral y expresión escrita. Los resultados se presentan en una escala de 10 a 160 puntos, con incrementos de 5 puntos, y su validez está respaldada por investigaciones

psicométricas que han demostrado su confiabilidad y precisión en la medición del dominio del idioma inglés (Duolingo, 2023). Desde un enfoque psicométrico, el DET cumple con los estándares de calidad establecidos por la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME), garantizando que la prueba es válida y confiable para la evaluación del nivel de inglés de los candidatos. Además, estudios recientes han demostrado que la estructura adaptativa del DET mejora la precisión de la medición y reduce la fatiga del examinado, lo que lo hace comparable con otras pruebas estandarizadas como el TOEFL y el IELTS (Duolingo, 2023).

- **ENLACE (México, hasta 2013) y PLANEA (Programa Nacional de Evaluación de los Aprendizajes, México):** Estas pruebas fueron diseñadas por el **Instituto Nacional para la Evaluación de la Educación (INEE)** para evaluar el aprendizaje en educación básica y media superior en México. **ENLACE** se aplicó hasta 2013 y fue reemplazada por **PLANEA**, que mide habilidades en lectura y matemáticas en alumnos de tercero de secundaria y último año de educación media superior (INEE, 2019). Ambas pruebas han sido validadas en términos de **validez de contenido** mediante expertos en educación y alineación con los planes de estudio nacionales. La confiabilidad de los ítems ha sido evaluada con el coeficiente **alfa de Cronbach**, mostrando valores superiores a **0.85** en la mayoría de sus secciones. Además, PLANEA ha incorporado modelos de **Teoría de Respuesta al Ítem (TRI)** para analizar la dificultad y discriminación de los ítems en diferentes niveles de desempeño (INEE, 2019). Estas pruebas permiten medir el logro académico de los estudiantes en relación con el currículo educativo y proporcionar información para la toma de decisiones en políticas educativas.
- **SABER (Colombia):** La prueba **SABER** es aplicada por el **Instituto Colombiano para la Evaluación de la Educación (ICFES)** y evalúa competencias desde la educación primaria hasta la educación superior. Existen diferentes versiones de la prueba, incluyendo **SABER 3°, 5°, 9°, 11° y PRO**, dependiendo del nivel educativo (ICFES, 2020). La prueba ha sido sometida a rigurosos análisis de **validez de contenido**, alineándola con los estándares curriculares nacionales. Su confiabilidad ha sido medida mediante coeficientes como el **alfa**

de Cronbach (≥ 0.90 en matemáticas y lectura). Además, **SABER 11 y SABER PRO** han implementado la **TRI**, permitiendo ajustar los ítems de acuerdo con el nivel de habilidad del estudiante y mejorar la precisión en la medición del desempeño (ICFES, 2020).

- **EVAU (Evaluación de Bachillerato para el Acceso a la Universidad, España):** La EVAU es una prueba de ingreso a la universidad en España. Evalúa conocimientos en diversas materias según la modalidad de bachillerato cursada por el estudiante (Ministerio de Educación y Formación Profesional, 2021). La prueba ha sido validada mediante estudios de **validez de contenido** que garantizan su alineación con los programas de estudio de bachillerato. La confiabilidad de los resultados se mide mediante la **consistencia interna** de los exámenes y pruebas piloto previas a la aplicación. En algunos casos, se han utilizado modelos de **TRI** para ajustar la dificultad de los ítems y mejorar la equidad de la evaluación (Ministerio de Educación y Formación Profesional, 2021).
- **AP (Advanced Placement, EE.UU.):** Los exámenes de **Advanced Placement (AP)** son diseñados por el **College Board** y permiten a estudiantes de secundaria en Estados Unidos obtener créditos universitarios en diversas materias. Cada examen evalúa conocimientos específicos y habilidades analíticas en áreas como matemáticas, ciencias, historia, lenguas y más (College Board, 2020). La validez de los exámenes AP se garantiza mediante paneles de expertos y análisis de **validez de contenido**. La confiabilidad de los exámenes se evalúa con coeficientes como el **alfa de Cronbach (> 0.85 en la mayoría de los casos)**. Además, los exámenes AP utilizan **Teoría de Respuesta al Ítem (TRI)** para ajustar la dificultad de los ítems y calcular puntajes equitativos entre distintos grupos de estudiantes (College Board, 2020).

2.7.2 Pruebas de desempeño

De acuerdo con Wiggins (1998), las pruebas de desempeño son instrumentos de evaluación que miden la capacidad de los individuos para aplicar conocimientos y habilidades en la resolución de problemas o en la ejecución de tareas en contextos reales o simulados. A diferencia de las pruebas tradicionales de opción múltiple, estas evaluaciones requieren que los evaluados realicen actividades complejas que reflejen situaciones del mundo real, permitiendo una medición más auténtica del aprendizaje y la competencia (Pellegrino et al. 2001). Según Darling-Hammond y Adamson (2014), este tipo de pruebas fomentan el

pensamiento crítico, la creatividad y la toma de decisiones, ya que los participantes deben demostrar su conocimiento a través de acciones concretas. Además, estas pruebas son ampliamente utilizadas en la educación, la formación profesional y la certificación laboral, con el objetivo de evaluar el nivel de competencia en diferentes áreas del saber (Brookhart S. M., 2013).

Tipos de Pruebas de Desempeño

Las pruebas de desempeño pueden clasificarse según su contexto de aplicación y su formato de evaluación:

1. Según el contexto de aplicación

- **Pruebas de desempeño en educación:** Estas pruebas evalúan las competencias de los estudiantes mediante actividades prácticas como ensayos, experimentos, estudios de caso o proyectos. Son utilizadas en programas como el *International Baccalaureate (IB)* y en evaluaciones de habilidades críticas como el *Collegiate Learning Assessment (CLA)* (Klein et al., 2007).
- **Pruebas de desempeño en certificaciones profesionales:** Son exámenes utilizados para obtener licencias o certificaciones en distintos campos profesionales. Ejemplos incluyen el *Objective Structured Clinical Examination (OSCE)* en medicina, el *Bar Exam* en derecho y el *Certified Public Accountant (CPA)* en contabilidad (Harden et al. 2016).
- **Pruebas de desempeño en el ámbito laboral:** Se aplican en procesos de selección y evaluación de empleados, utilizando técnicas como *Assessment Centers*, simulaciones y estudios de caso para medir habilidades profesionales (Arthur et al. 2003).

2. Según el formato de evaluación

- **Pruebas escritas de desempeño:** Involucran la producción de ensayos, estudios de caso y resolución de problemas complejos, donde se evalúa la argumentación y el análisis crítico (Brookhart S. M., 2013).
- **Pruebas prácticas:** Incluyen la ejecución de tareas específicas, como simulaciones clínicas, presentaciones orales, demostraciones técnicas o experimentos en laboratorios (Darling-Hammond, L., & Adamson, F., 2014).

2.7.3 Pruebas de conocimiento

Según Hyladina (2004), las pruebas de conocimiento son instrumentos de evaluación diseñados para medir el grado de adquisición, comprensión y retención de información en un área específica del saber. Su propósito principal es evaluar qué tanto los evaluados han aprendido sobre un tema determinado, generalmente en el ámbito educativo o profesional. Estas pruebas pueden ser utilizadas en diversos niveles, desde la educación primaria hasta la formación universitaria y profesional, y suelen aplicarse en contextos académicos, certificaciones y selección de personal.

Según Nitko y Brookhart (2011), las pruebas de conocimiento presentan las siguientes características:

- **Evaluación del saber declarativo:** Se centran en medir qué tanto el evaluado conoce sobre un tema específico.
- **Uso de formatos estructurados:** Generalmente incluyen preguntas de opción múltiple, verdadero/falso, respuesta corta y ensayo.
- **Aplicabilidad en distintos contextos:** Son utilizadas en educación, certificaciones profesionales y selección de personal.
- **Énfasis en la memorización y comprensión:** Pueden evaluar desde conocimientos básicos hasta la aplicación de conceptos en problemas complejos.

Las pruebas de conocimiento pueden clasificarse según su propósito y estructura:

Según su propósito:

- **Pruebas diagnósticas:** Se aplican antes de la instrucción para determinar el nivel de conocimientos previos (Pellegrino et al. 2001).
- **Pruebas formativas:** Se utilizan durante el proceso de aprendizaje para monitorear el progreso del estudiante (Black, P., & Wiliam, D., 1998).
- **Pruebas sumativas:** Se aplican al final de un curso o programa de formación para evaluar el nivel de logro alcanzado (Harlen, 2007).

Según su formato:

- **Pruebas de opción múltiple:** Evalúan conocimientos mediante la selección de respuestas correctas entre varias alternativas (Haladyna et al. 2002).

- **Pruebas de respuesta corta o ensayo:** Requieren que los evaluados expliquen o desarrollen sus respuestas, demostrando mayor profundidad en la comprensión del contenido.
- **Pruebas de Completación y verdadero/falso:** Miden el reconocimiento de información correcta o incorrecta sobre un tema.

Las pruebas de conocimiento son utilizadas en distintos contextos, desde la educación hasta el ámbito laboral y profesional:

- **En educación:**
 - **Exámenes escolares y universitarios:** Evaluaciones aplicadas en instituciones educativas para medir el aprendizaje en diversas asignaturas.
 - **Exámenes de admisión (ej., SAT, GRE, EXANI-II):** Pruebas que evalúan conocimientos generales para el ingreso a la educación superior.
- **En certificaciones profesionales:**
 - **Exámenes de certificación profesional (ej., CPA, Bar Exam, PMP):** Evalúan conocimientos en áreas específicas como contabilidad, derecho o gestión de proyectos.
- **En el ámbito laboral:**
 - **Pruebas de conocimientos técnicos en procesos de selección:** Se utilizan para medir el dominio de habilidades y conocimientos en áreas específicas, como tecnología, ingeniería o ciencias de la salud.

Las pruebas de conocimiento presentan beneficios y limitaciones que deben ser consideradas en su diseño y aplicación:

- Son fáciles de administrar y calificar, especialmente cuando se utilizan formatos de opción múltiple (Nitko, 2011).
- Permiten evaluar una gran cantidad de contenido en un tiempo reducido.
- Proporcionan información objetiva sobre el nivel de conocimiento de los evaluados.

Los desafíos que enfrenta son:

- Pueden fomentar el aprendizaje memorístico en lugar del desarrollo de habilidades analíticas y críticas (Harlen, 2007).
- Las pruebas de opción múltiple pueden ser afectadas por el azar si los evaluados responden al azar (Haladyna et al., 2002).
- La validez y confiabilidad de estas pruebas dependen de un diseño cuidadoso y de la claridad en la formulación de los ítems.

Para garantizar que una prueba cumpla con su propósito de evaluación de manera objetiva y precisa, no basta con su adecuada construcción; es necesario someterla a un análisis riguroso basado en principios psicométricos. La psicometría proporciona los métodos y herramientas para evaluar la validez, confiabilidad y estructura interna de un test, asegurando que sus puntajes reflejen con precisión el constructo que se pretende medir (Muñiz, 2018; AERA, APA, & NCME, 2014).

2.8 Psicometría

El termino Test psicométrico fue utilizado por primera vez en el año de 1890 por James Cattell quien lo utilizo para hacer referencias a una serie de pruebas psicológicas que se encargaban de medir diferencias individuales y de esa forma este concepto Test Psicométrico fue empleado en las aplicaciones de otras pruebas con las que se buscaba un nuevo método de medición Psicológica (Cortada de Kohan, 2002).

Melia (1990) define la psicometría como la teoría y metodología acerca de la medida y sus técnicas en psicología, implicando sus condiciones, limitaciones y finalidades. La psicometría se refiere al campo de la teoría y práctica sobre la elaboración, evaluación y aplicación de las medidas en psicología. Mientras que Muñiz (2018) define la Psicometría como "... el conjunto de métodos, técnicas y teorías implicadas en la medición de variables psicológicas... lo específico de la Psicometría sería su énfasis y especialización en aquellas propiedades métricas exigibles a las mediciones psicológicas independientemente del campo sustantivo de aplicación y de los instrumentos utilizados" (citado por (Ortiz, 2014).

En 1905, los franceses Binet y Simón crearon el primer test psicométrico el cual fue, la Escala Métrica de la Inteligencia con la que se introdujo en psicología el concepto de edad

mental, seguido vinieron más estudios, y así se agregó el concepto coeficiente intelectual o CI, gracias al psicólogo Alemán William Stern en 1912. En paralelo Robert S. Wadsworth en 1916, creo el primer test de personalidad denominado, Cuestionario de Datos de Personalidad, reconocido como el primer test de personalidad (Aliaga Tovar, 2007).

Para los próximos años, los test fueron contruidos basándose en el enfoque de la teoría clásica de los test, creada por el psicólogo inglés, Charles Spearman en la segunda década del siglo XX; en las tres últimas décadas de ese mismo siglo, desarrollaron otro enfoque para la reconstrucción de test que fue llamado Teoría de Respuesta al Ítem (TRI), sin embargo, hay muy pocos test basados en esta teoría (Aliaga Tovar, 2007).

La psicometría no solo permite evaluar la calidad de un instrumento de medición, sino que también proporciona criterios fundamentales para determinar su capacidad de medir adecuadamente el constructo de interés. Dentro de este proceso, la validez juega un papel central, ya que garantiza que la prueba realmente evalúe lo que se propone medir. En este sentido, la validez de contenido es una de las primeras evidencias a considerar, ya que se enfoca en la representatividad y relevancia de los ítems en relación con el dominio teórico que se pretende medir (AERA, APA, & NCME, 2014; Muñiz, 2018).

2.8.1 Validez

Gómez et al. (2009) establece que la validación es un proceso continuo, que incluye procedimientos diferentes para comprobar si el cuestionario mide realmente lo que dice medir, planteado de otra manera tiene que ver con las decisiones que se tomen a partir de las puntuaciones obtenidas en la prueba. Mientras que para Soriano Rodriguez (2014) las principales propiedades de una medición son la confiabilidad y la validez. Un instrumento puede ser confiable, pero no necesariamente valido para una población en particular, o en el peor de los casos, que el instrumento haya sido manipulado para obtener ciertos resultados. En lugar de clasificar los tipos de validez, Messick (1989) propone recolectar diferentes tipos de evidencias con base al uso y objetivos del instrumento, entre ellas la evidencia del contenido, del constructo y su valor predictivo. Además, debe tenerse en cuenta que la validez no es una propiedad intrínseca de los instrumentos, sino que dependerá del objetivo de la medición, la población y el contexto de aplicación, por lo que un instrumento puede ser válido para un grupo en particular, pero no para otros (Soriano Rodriguez, 2014).

2.8.2 Validez de Contenido

La validez de contenido es un juicio lógico que pretende determinar si los ítems o preguntas propuestas reflejan el dominio de contenido (conocimientos, habilidades o destrezas) que se desea medir, es por ello que este tipo de validez es totalmente relevante para la construcción de pruebas, esta tipología de validez también puede considerar criterios internos y externos.

Entre la validez interna y externa encontramos una serie de criterios:

validez interna: busca que haya calidad en sus contenidos, cobertura del contenido, verificar si hay complejidad cognitiva, uso adecuación lingüística en el test, habilidades, entre **validez externa:** Esta se enfoca en aspectos como la transferencia y generación de conocimientos, finalmente la sensibilidad de la instrucción velando porque esta sea de claridad y con alto nivel de comprensión (Shuttleworth, 2020).

La validez de contenido se puede establecer en distintas situaciones, sin embargo, mencionaremos dos ya que son las en las que se utiliza con mayor frecuencia, 1) el diseño de una prueba. 2) en la validación de un instrumento construido para una población distinta, afianzada en estos dos escenarios podemos decir que este tipo de validez se enfoca en la concordancia de los contenidos en relación a los ítems y con ello alcanzar una medición relevante y representativa del constructo según Urrutia Egaña et al. (2014).

2.8.3 Metodología para validar contenidos

Existen diversos métodos para evaluar la validez de contenido, los cuales buscan garantizar que los ítems de un instrumento representen adecuadamente el constructo de interés. Entre estos métodos, el juicio de expertos es una de las estrategias más utilizadas, ya que permite evaluar la claridad, relevancia y representatividad de los ítems a través de la opinión de especialistas en el área. Este proceso proporciona evidencia fundamentada sobre la adecuación del contenido de la prueba antes de su aplicación (AERA, APA, & NCME, 2014; Hernández-Nieto, 2002; Muñiz, 2018).

2.8.4 Validez de Contenido Mediante Juicio de Expertos

Uno de los métodos más populares para darle validez de contenido a una prueba, es teniendo en cuenta el procedimiento conocido como “**validez de contenido por expertos**”, el cual consiste en la revisión por parte de los expertos para que la prueba o test apruebe ciertos requisitos y se tenga mayor seguridad de que el test aplicado en investigación sea el

adecuado, y que contenga los temas que se requieren ver durante la revisión de un tema de estudio (Sanchez, 2021).

Autores como Urrutia Egaña et al. (2014), en su artículo **“Métodos Óptimos para Determinar Validez de Contenido”** proponen fases para llevar a cabo el proceso de validez de contenido mediante expertos:

- a) Realizar una definición del universo de observaciones admisibles.
- b) Determinar quiénes son los expertos en el universo.
- c) Exponer (por parte de los expertos) el juicio por medio de un procedimiento concreto y estructurado sobre la validez del contenido.
- d) Elaborar un documento donde se resumen los datos recopilados anteriormente

Estas fases en conjunto muestran un desarrollo eficiente en la construcción de juicios óptimos de la validez de contenidos realizada por expertos.

Autores como Skjong y Wentworht (2000), han propuesto una serie de lineamientos para desarrollar “Validez de contenido mediante juicio de Expertos” entre ellos encontramos:

- (a) Construir plantillas con sus instrucciones.
- (b) Llevar a cabo la selección de experto con conocimientos del área y brindarles el entrenamiento pertinente.
- (c) Exponer debidamente el contexto.
- (d) Viabilizar la discusión y el acuerdo entre expertos.

Es de total importancia que las instrucciones facilitadas a los jueces o expertos sean claras y precisas en cuanto a las dimensiones e indicadores que mide cada reactivo, de esa manera el experto tendrá mayor conocimiento al momento de validar cada ítem.

Las autoras Escobar Pérez & Cuervo Martínez (2008) en su artículo **“Validez de contenido y juicio de expertos: una aproximación a su utilización”**, proponen también una serie de pasos detallados sobre el procedimiento genérico de cómo realizar el proceso para desarrollar un juicio de expertos claro y preciso:

1. Definir el objetivo del juicio de expertos:

Se debe mostrar de forma clara cual es finalidad de este procedimiento ya que en el transcurso nos damos cuenta que la validez de contenido mediante juicio de expertos puede ser utilizado en tres aspectos como:

- (a) Establecer la equivalencia semántica de una prueba que se encuentra validada en otro idioma.
- (b) Evaluar la adaptación cultural, es decir, el objetivo de los jueces es evaluar si los ítems de la prueba miden el mismo constructo en una cultura distinta.
- (c) Validar contenido en una prueba diseñada por un grupo de investigadores.

2. La selección de los jueces:

Para llevar a cabo este paso se deben considerar una serie de criterios, estos pueden ser dados o estructurados por el especialista, pero podríamos mencionar algunos de los más comunes, Experiencia académica o en el campo y formación profesional.

3. Explicar cada una de las dimensiones:

Como los indicadores que es lo que se está midiendo con ellos, esto permitirá que los expertos puedan evaluar 3 criterios esenciales como lo son la relevancia, la suficiencia y la pertinencia del ítem.

4. Manifiestar a los expertos cual es el objetivo de la prueba.

5. La plantilla a utilizarse debe ser diseñada de acuerdo con los objetivos propuestos en la evaluación.

6. Calcular la concordancia entre jueces:

Para ello podemos hacer uso de estadísticos como Kappa y Kendall.

7. Para finalizar se deben elaborar las conclusiones del juicio, ya que estas serán utilizadas para la descripción psicométrica de la prueba (Escobar Pérez & Cuervo Martínez, 2008).

El autor Rubén Sánchez en su artículo “**El tema de validez de contenido en la educación y la propuesta de Hernández-Nieto**” destaca que es una herramienta clave para la evaluación de la validez de contenido a través de juicio de expertos, proporcionando un enfoque cuantitativo y estructurado que permite determinar la concordancia entre evaluadores. Su principal aporte es el desarrollo del Coeficiente de Validez de Contenido (CVC), un índice

que mide la pertinencia, claridad y representatividad de los ítems en función de las valoraciones otorgadas por un panel de jueces especializados (Hernández-Nieto, 2002). Este método se destaca por su capacidad para ofrecer una estimación objetiva de la validez, minimizando la subjetividad inherente al proceso de revisión experta (Sanchez, 2021). Además, establece criterios específicos de evaluación, como el uso de escalas Likert y el análisis de niveles de dificultad, lo que permite obtener una medición precisa y replicable de la calidad del instrumento evaluado (Sanchez, 2021). Gracias a su aplicabilidad en contextos educativos y psicológicos, el CVC se ha convertido en una referencia metodológica fundamental en la construcción y validación de pruebas de medición.

(Hernandez-Nieto, 2002) recomienda la participación de entre tres y cinco expertos respecto a cada uno de los diferentes ítems y al instrumento en general. Para ello, tras la aplicación de una escala tipo Likert de cinco alternativas, se calcula la media obtenida en cada uno de los ítems y, en base a esta, se calcula el CVC para cada elemento así:

$$CVC_i = \frac{M_x}{V_{m\acute{a}x}}$$

Donde M_x representa la media del elemento en la puntuación dada por los expertos y $V_{m\acute{a}x}$ la puntuación máxima que el ítem podría alcanzar. Por otro lado, debe calcularse el error asignado a cada ítem (Pe_i), de este modo se reduce el posible sesgo introducido por alguno de los jueces, obtenido mediante:

$$Pe_i = \left(\frac{1}{j}\right)^j$$

siendo j el número de expertos participantes. Finalmente, el CVC se calcularía aplicando $CVC = CVC_i - Pe_i$. Respecto a su interpretación Hernandez-Nieto (2002) recomienda mantener únicamente aquellos ítems con un CVC superior a 0.80.

Si bien la validez de contenido garantiza la pertinencia de los ítems mediante el juicio de expertos, es necesario complementarla con métodos empíricos. En este sentido, el análisis factorial permite evaluar la estructura interna del instrumento, verificando si los ítems se agrupan de manera coherente en las dimensiones teóricas propuestas (AERA, APA, & NCME, 2014; Muñiz, 2018).

2.8.5 Método basado en el Análisis Factorial

El análisis factorial exploratorio es un conjunto de métodos multivariados de interdependencia, cuya finalidad es identificar una estructura de variables subyacentes inobservables a un conjunto amplio de datos observables (Pérez, 2010).

El análisis factorial tiene sus inicios en los trabajos de Galton en 1889, quien planteó la idea de un rasgo latente. Galton propuso que, si dos variables están correlacionadas entre sí, es debido a que poseen un rasgo latente en común, que explica parte de la varianza de las variables. En 1901 Pearson desarrolló el coeficiente de correlación y aportó principios fundamentales al análisis. Sin embargo, el creador por excelencia del análisis factorial fue Spearman en 1904. Él estudió las correlaciones en diferentes pruebas de aptitudes, las cuales según Pérez de Pérez (2006) reflejaban un factor subyacente de inteligencia. Más adelante, a mitad de siglo XX, Thurstone consolidó las bases metodológicas y reformuló el análisis.

El análisis factorial se divide en Análisis Factorial Exploratorio (AFE) y Análisis Factorial Confirmatorio (AFC) (Burga Leon, 2006). El análisis factorial exploratorio es un método estadístico que encuentra patrones de asociación entre ítems y variables para establecer una estructura de dimensiones en base a las correlaciones entre las variables observadas (Martínez Arias et al., 2006; Rojas Torres, 2014). Este se utiliza principalmente en las etapas iniciales de desarrollo de test ya que no se imponen restricciones al modelo (Ledesma et al., 2019) En cambio, el análisis factorial confirmatorio comprueba el ajuste de los ítems conforme a un modelo teórico previamente establecido (Rojas, 2014). En este, las restricciones son más rigurosas, definiendo el número de factores y su relación entre ellas (Ferrando P. J.-C., 2010).

En el contexto de la validez de contenido, el análisis factorial ayuda a determinar si los ítems de una prueba están alineados con las dimensiones teóricas del constructo que se está evaluando. Al realizar un análisis factorial, se pueden identificar los factores que subyacen a los ítems y evaluar la pertenencia de cada ítem a un factor específico. Esto proporciona evidencia cuantitativa sobre la validez de contenido, asegurando que los ítems son representativos y relevantes para el constructo en cuestión (Pedrosa et al, 2014).

El método planteado por Tucker se basa en el análisis factorial de las puntuaciones otorgadas por los expertos en cuanto a la relevancia de los ítems, pudiendo obtener dos factores diferenciados. El primero de ellos, puede interpretarse como una adecuación muestral de los ítems para constituir un test, al considerar el test como una muestra representativa de la

variable de interés. Por otra parte, el segundo permite detectar las diferencias de puntuaciones dadas en la evaluación de los expertos (Amaya L. J. L. and Pacheco D., 2002).

2.8.5.1 Supuestos del análisis factorial

De acuerdo con Pérez (2010) el análisis factorial trabaja en base a ciertos supuestos que al seguirlos se podrá obtener mediciones significativas:

- Normalidad: los datos siguen una distribución normal.
- Linealidad: la relación entre las variables debe ser lineal.
- Multicolinealidad: Los ítems no deben tener correlaciones elevadas o redundantes

2.8.5.2 Fases del Análisis Factorial

El análisis factorial exploratorio consiste en cuatro fases principales 1. Preparación inicial de datos, 2. extracción de factores, 3. rotación de factores e 4. interpretación de factores.

1. Preparación inicial de datos:

Es preciso realizar un análisis previo para depuración de errores, verificación de existencia de puntuaciones marginales, comprobación de supuestos y determinación de la matriz de correlaciones a utilizar (Martínez Arias et al., 2006).

El análisis puede verse afectado por la existencia de outliers o puntuaciones marginales, ya que el AFE se basa en las correlaciones entre variables, las cuales son estimadas mediante la media, que es una medida de tendencia central muy sensible a los valores extremos. Por tal razón, conviene en primer lugar realizar un análisis inicial de datos para detectar dichos valores extremos. Existen diferentes formas de realizar el análisis. Comúnmente, se calculan las puntuaciones típicas de cada variable y se identifican las puntuaciones que se encuentren fuera del rango ± 3 . También se puede visualizar los gráficos de caja y bigotes, en los cuales se muestran claramente los valores atípicos como puntos aislados en los extremos de los bigotes. Otra alternativa es utilizar el procedimiento de distancia Mahalanobis (D^2) para identificar los casos atípicos que superen la significación de $p < .001$ (Pérez, 2010).

Otro análisis previo esencial es evaluar si los ítems están suficientemente interrelacionados para ser analizados por medio del AFE. Para esto, se puede realizar algunas pruebas estadísticas llamadas medidas de adecuación muestral, como el test de esfericidad de Bartlett, el cual pone a prueba la hipótesis nula de que las variables no están correlacionadas en la población. Si los resultados son significativos ($p < .05$) se rechaza la hipótesis nula y se considera que las variables están suficientemente correlacionadas para realizar el AFE

(Ferrando P. J.-C., 2010). También se puede realizar la medida de adecuación muestral de Kaiser-Mayer-Olkin (KMO), este evalúa hasta qué punto las puntuaciones en cada una de las variables son predecibles desde las demás. El rango de valores es de 0 a 1, si $KMO > .90$ el test es muy bueno, es notable si $KMO > 0.80$ y medianamente aceptable si $KMO > 0.70$ valores menores que 0.60 sugerirán interrelaciones bajas entre los ítems (Pérez, 2010).

Habiendo realizado estos análisis iniciales, se procede a verificar los supuestos de normalidad, linealidad y multicolinealidad. Usualmente se utiliza pruebas de contraste de bondad de ajuste como Shapiro-Wilk y Kolmogórov-Smirnov para evaluar la distribución normal. Sin embargo, también es recomendable estimar los índices de asimetría y curtosis, considerando que los valores dentro del rango ± 1.5 indican variaciones leves de la normal y, por lo tanto, los datos son adecuados para realizar el AFE. También se puede visualizar los gráficos Q-q plot, para determinar si los datos se ajustan a una distribución normal (Pérez, 2010).

El supuesto de linealidad es evaluado mediante diagramas matriciales de dispersión. Los puntos deben organizarse a lo largo de una línea recta para que exista linealidad. Además, se puede realizar una Estimación Curvilínea por medio de análisis de regresión múltiple (Pérez, 2010).

Por último, el supuesto de colinealidad se analiza por medio de las correlaciones entre variables, si estas son superiores a 0.90 existe multicolinealidad y terminará afectando la estabilidad de la solución factorial. También puede estimarse los índices de tolerancia y la inflación de la varianza (VIF). Valores pequeños de tolerancia (menores que 0.10) y elevados de VIF (mayores que 10) demuestran alta colinealidad (Pérez, 2010).

Ahora bien, se debe decidir el tipo de matriz de correlaciones a utilizar, esta decisión debe ir en concordancia con el tipo de datos. Usualmente se utiliza la matriz de correlaciones de Pearson, sin embargo, surge un problema al analizar ítems dicotómicos, ya que este tipo de correlación asume que las variables están medidas a nivel de intervalo y tienen distribución normal. Para una estimación más precisa, la solución propuesta es utilizar una matriz de correlaciones tetracóricas, estas son estimaciones máximo-verosímil de la correlación de Pearson que había en las variables antes de ser dicotomizadas, asumiendo que tras esta dicotomía se encuentran dos variables distribuidas normalmente (Burga León, 2006; Ferrando P. J., 1966).

2. Extracción de factores:

En esta fase se determina el número de factores y el método de extracción. Tras haberse verificado los supuestos, se procede a seleccionar un método de extracción de factores para estimar una solución factorial inicial. Los más utilizados en el AFE son el de Componentes Principales, Ejes Principales y Máxima Verosimilitud (Mavrou, 2015; Pérez, 2010). El método de Componentes Principales explica la mayor cantidad de varianza en los datos observados (Mavrou, 2015). Este es un método de extracción de factores, pero en realidad, no es un procedimiento para estimar un modelo factorial, ya que funciona en base a la varianza total, por lo que no siempre produce los mismos resultados que un modelo factorial (que sí distingue entre la varianza común y la varianza de error). Por tal razón no es un método apropiado a considerar, sin embargo, es mencionado aquí debido a que existe gran confusión entre el análisis factorial y componentes principales y muchas veces es utilizado injustificadamente llevando a mediciones erróneas (Pérez, 2010).

Otros métodos de estimación son los basados en Mínimos Cuadrados Ordinarios y Máxima Verosimilitud. Los estimadores por Mínimos Cuadrados buscan determinar una solución factorial que minimice la suma de cuadrados de las diferencias entre las correlaciones observadas y las reproducidas desde el modelo, logrando que los residuales sean lo más cercanos a 0 posible (Ferrando P. J.-C., 2010). Un método relevante dentro de este criterio es el de Ejes Principales, el cual es un método iterativo basado en la extracción de comunalidades de la varianza, este suele ser recomendado cuando los datos violan el supuesto de normalidad. Tiene la ventaja de ofrecer soluciones estables y recuperar valores de correlaciones débiles en muestras pequeñas. Sin embargo, no proporciona índices de ajuste ni intervalos de confianza (Pérez, 2010). También existe otro método de Máxima Verosimilitud, que permite estimar la significación estadística de los pesos factoriales y su ajuste al modelo (Ferrando P. J.-C., 2010). Sin embargo, este es un método estadístico inferencial que presupone que las variables son continuas y distribuidas con normalidad, en ausencia de casos atípicos (Mavrou, 2015). Por lo cual es una buena elección cuando los datos se distribuyen con normalidad (Pérez, 2010). En contraste, el incumplimiento de los supuestos conllevará a resultados problemáticos y no fiables.

El siguiente aspecto para considerar es el número de factores a extraer, lo cual debe fundamentarse tanto en la teoría como en la evidencia empírica (Mavrou, 2015). En la literatura científica no existe un consenso claro sobre un criterio para verificar el supuesto

de unidimensionalidad. El criterio más utilizado es la regla de Gutman-Kaiser, en el cual se debe conservar aquellos factores con valores mayores que 1 afirma Martínez Arias et al. (2006). Según el criterio de Gutman-Kaiser, para que el test sea aceptado como unidimensional debe existir un solo autovalor superior a uno asume Aune et al. (2019). Esta regla implica que un factor para que sea útil debería explicar tanta varianza como una variable individual (Pérez, 2010). Sin embargo, presenta algunos problemas ya que se basa en los valores propios de la matriz sin reducir que son equivalentes a la proporción de la varianza total, cuando lo que realmente interesa es la varianza común (Ferrando P. J.-C., 2010). Un criterio clave para la extracción de factores analizar es el porcentaje de varianza explicada por la solución factorial obtenida (Pérez, 2010). Un test se considera unidimensional si explica una buena proporción de la varianza total y si los sucesivos factores explican en mucho menor cantidad entonces se considera unidimensional. Existen diversas opiniones sobre la cantidad de varianza explicada del primer factor. En general, el primer factor debería explicar entre el 30 a 40%, de la varianza (Burga León, 2006; Ferrando P. J., 1966).

Además, se utiliza el test de sedimentación, un gráfico grafico bivariado de los autovalores de la matriz de correlación original sin rotar (Pérez, 2010). Cuando entre el primer autovalor se identifica un salto brusco en la pendiente y los restantes comienzan a generar una pendiente de poca inclinación se acepta la unidimensionalidad según Aune et al. (2019). Sin embargo, este método presenta dos problemas, el primero es que se apoya en la subjetividad de la inspección visual y aunado a esto, se basa en la varianza total, sin distinguir la varianza común del error. Una solución es utilizar las comunales en lugar de los valores propios de la matriz sin reducir (Ferrando P. J.-C., 2010).

Alternativamente, Ferrando y Anguiano-Carrasco (2010) proponen evaluar las correlaciones residuales como un método más acertado para determinar el número de factores. Si la distribución de frecuencias residuales es normal y centrada en torno a una media de cero indicará que el número de factores es adecuado.

Existen otras evidencias para evaluar la unidimensionalidad, mediante el Análisis Factorial Confirmatorio, en el cual se impone una estructura unifactorial. También mediante el Test Estadístico de Unidimensionalidad Esencial de Stout, sin embargo, solo puede ser aplicado en muestras de por lo menos 500 participantes según Aune et al. (2019).

3. Rotación de factores:

En esta fase se elige el procedimiento de rotación de los factores. El resultado inicial del análisis factorial requiere un procedimiento de rotación para facilitar su interpretación. Sin embargo, para test unidimensionales la rotación no es tan imprescindible, ya que se puede interpretar directamente (Ferrando P. J.-C., 2010). Existen varios métodos de rotación de factores, comúnmente se clasifican en rotaciones ortogonales (varimax, quartimax y equimax) u oblicuas (oblimin direct, promax, orthoblique, procrustes) (Mavrou, 2015).

Los métodos más comunes son Varimax y Promax. Las rotaciones oblicuas como la Promax son más apropiadas para variables psicológicas, ya que los factores normalmente están correlacionados, estas permiten obtener soluciones precisas, simples y realistas. Las rotaciones ortogonales son consideradas más simples y fáciles de interpretar, pero asumen que los factores no están correlacionados, normalmente se utilizan cuando las correlaciones son menores que 0.32 (Pérez, 2010).

4. Interpretación de factores:

La tarea final del análisis es interpretar y denominar los factores. Para esto, se analiza en la matriz rotada, los patrones de correlaciones altas y bajas en los distintos factores posibles (Pérez, 2010). Se asume que los ítems pertenecientes a una misma dimensión teórica deben presentar altas correlaciones entre sí (Rojas, 2014).

En las Ciencias Sociales, particularmente en la medición de constructos psicológicos las saturaciones generalmente son moderadas (menores que 0.32), por tanto, saturaciones por encima de .50 son consideradas saturaciones fuertes. Se recomienda que la dimensión debe poseer al menos 4 ítems con correlaciones superiores a 0.40 (Mavrou, 2015). En la aplicación del AFE, una exigencia muy importante es que el tamaño de la muestra debe ser grande para asegurar un menor error de muestreo. Para ítems dicotómicos, se recomienda un numero alrededor de 300 sujetos en las muestras.

Por último, es muy importante la estimación de índices de bondad de ajuste, como el GFI (Goodness of fit index), el cual indica la proporción de covarianza explicada por el modelo. Valores superiores a 0.95 indican un buen ajuste (Lloret-Segura, 2014). También está el SRMR (Standardized Root Mean Squared Residual) el cual informa sobre la diferencia estandarizada entre las correlaciones observadas y las predictivas. Valores inferiores a 0.08 se consideran un buen ajuste. En general, valores menores que 0.10 son considerados aceptables (Cangur, 2015).

El análisis factorial permite evaluar la estructura interna de una prueba, identificando cómo se agrupan los ítems en factores y asegurando que midan el constructo de manera coherente. Sin embargo, para que un instrumento sea realmente útil en la medición, no solo debe ser válido, sino también confiable. La confiabilidad se refiere a la consistencia de los puntajes obtenidos en diferentes aplicaciones del test, asegurando que las mediciones sean precisas y reproducibles. Para ello, se emplean métodos como el Alfa de Cronbach y el Omega de McDonald, los cuales permiten estimar la estabilidad y homogeneidad de los ítems en la prueba (AERA, APA, & NCME, 2014; Muñiz, 2018).

2.9 Fiabilidad

La fiabilidad es un criterio de la calidad métrica que evalúa el grado de precisión de la medida obtenida en un test (Navas Ara M. J., 2002). Las interpretaciones de las puntuaciones en un test son fiables cuando las medidas son consistentes y carecen de errores de medida (Muñiz, 2018). El análisis de fiabilidad proporciona información sobre la cantidad de error aleatorio que contiene la medida. Es decir, factores no relacionados al constructo medido, pero que influyen en la puntuación del individuo. Estos errores pueden emerger de dos fuentes principales: a) fluctuaciones o inconsistencias de las puntuaciones a lo largo del tiempo y b) diferencias en el contenido entre los ítems o incongruencias entre las respuestas del test. A partir de estas fuentes de error, se conceptualiza la fiabilidad en primer lugar, como estabilidad temporal y, en segundo lugar, como consistencia interna (Navas, 2001).

El coeficiente más utilizado para estimar la fiabilidad es el coeficiente alfa propuesto por Cronbach en 1951 (Cronbach, 1951). Este coeficiente se utiliza para evaluar la consistencia interna de los instrumentos psicométricos, ya que permite estimar el grado de correlación entre los ítems que componen una escala. Su valor varía entre 0 y 1, y se considera que valores iguales o superiores a .70 son aceptables para fines exploratorios. Por el contrario, valores bajos pueden reflejar escasa homogeneidad entre los ítems, problemas en la redacción o la posible multidimensionalidad del instrumento. Un valor particularmente bajo, como $\alpha \leq .50$, sugiere que los ítems no presentan una correlación suficiente para conformar una escala confiable. Por ello, es fundamental reportar e interpretar de manera adecuada este índice al analizar la validez y confiabilidad de los instrumentos aplicados (Oviedo & Campo-Arias, 2005; Viladrich, Angulo-Brunet, & Doval, 2017).

Se representa de la siguiente manera.

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_i^2}{\sigma^2} \right)$$

donde:

- α representa el coeficiente alfa de Cronbach.
- k es el número de ítems en la escala.
- $\sum \sigma_i^2$ es la suma de las varianzas individuales de cada ítem.
- σ^2 es la varianza total del test.

Existen algunos factores que influyen en la estimación del α de Cronbach, como la longitud del test, de forma que cuantos más ítems tenga el test, el valor del coeficiente sería más alto. Además, la fiabilidad depende de las características de la muestra a la que se aplicó el test, esta puede aumentar cuando la muestra es heterogénea, ya que, a mayor variabilidad, existirá mayor correlación. La muestra también debería ser numerosa, por lo que desde este enfoque la fiabilidad es una propiedad de las puntuaciones del test en una muestra determinada (Navas, 2001). Es importante señalar que los coeficientes de consistencia interna apoyan el supuesto de unidimensionalidad (Morales Vallejo, 2007). Debido a que se espera que los ítems que conforman un test midan el mismo constructo, cuando los datos son multidimensionales el valor de consistencia interna será bajo (Frias-Navarro, 2022). Por tanto, se recomienda utilizar estos análisis cuando el test es unidimensional o con contenido homogéneo, sin embargo, no es apropiado en test de velocidad (Navas, 2001).

En la actualidad se acumula evidencia sobre algunas limitaciones del alfa de Cronbach. Primeramente, el **alfa de Cronbach** asume que los ítems miden un mismo constructo y que están correlacionados entre sí. También supone que las respuestas son politómicas en escala ordinal o continuas. Además, la independencia o no correlación de los errores. Si el supuesto de la tau-equivalencia no se cumple, el valor del alfa es infraestimado, por el contrario, si el supuesto de no correlación entre los errores se viola, el valor del alfa se sobreestima (Frias-Navarro, 2022), aunado a esto el alfa de Cronbach también depende del número de alternativas de respuesta (Ventura-Leon, 2017).

El omega de McDonald Ω

Este surgió como un método alternativo debido al uso inadecuado del alfa de Cronbach cuando no se cumplen sus supuestos. Este, al igual que el alfa de Cronbach, está basado en el promedio de las correlaciones entre los ítems. El omega de McDonald no exige que los errores estén correlacionados ni la equivalencia.

La principal diferencia entre el alfa de Cronbach y el omega de McDonald es que éste se fundamenta en la matriz de correlación Policóricas y entre los ítems, en lugar de la matriz de covarianza o correlación de Pearson, y por ello es más apropiado cuando los datos son ordinales o dicotómicos (Frias-Navarro, 2022). Además, el coeficiente omega de McDonald trabaja con las cargas factoriales estandarizadas, por lo que no depende del número de ítems (Ventura-Leon, 2017).

Los valores del coeficiente de fiabilidad oscilan entre 0 a 1, su interpretación depende del uso previsto para el instrumento. Si es para la toma de decisiones debe ser mayor que 0.85, si es para investigación puede ser mayor que 0.60, si es para descripción de grupos puede ser mayor que 0.60 (Morales Vallejo, 2007). En general, valores menores que 0.70 indicarán baja fiabilidad y valores mayores que 0.95 serían considerados como indicadores de redundancia o duplicación de ítems (Frias-Navarro, 2022). La fiabilidad juega un rol trascendental en el uso de test con fines de investigación, una baja fiabilidad limitaría el valor de los análisis estadísticos, asimismo, en la utilización de test para la toma de decisiones, una baja fiabilidad afectaría una toma de decisión acertada (Frias-Navarro, 2022). Un coeficiente de fiabilidad bajo indica que hay errores de medición, puede ser producto de una inadecuada formulación de los ítems, causando confusión en los individuos, también puede deberse a que la muestra es muy homogénea (Morales Vallejo, 2007).

La **fiabilidad** de una prueba es un requisito fundamental para garantizar la precisión y estabilidad de las mediciones obtenidas. Sin embargo, la forma en que se conceptualiza y analiza la fiabilidad varía según el marco teórico de referencia en la medición psicométrica. En este contexto, emergen dos enfoques predominantes: la **Teoría Clásica de los Tests (TCT)** y la **Teoría de Respuesta al Ítem (TRI)**. La **TCT** asume que los errores de medición son aleatorios y que la confiabilidad puede estimarse a partir de la consistencia interna o la estabilidad temporal del instrumento, mientras que la **TRI** modela la relación entre las respuestas de los examinados y las características de los ítems, permitiendo una evaluación más precisa de la fiabilidad en función del nivel de habilidad del evaluado. La selección

entre estos enfoques depende del propósito de la medición y del nivel de precisión requerido en la interpretación de los puntajes (Muñiz, 2018; AERA, APA, & NCME, 2014).

2.10 Teorías de Medición

Ahora bien, la ciencia que permite la medición de fenómenos psicológicos es la psicometría, ésta se fundamenta en modelos de medición, basados en la estadística y la probabilidad, para poder establecer relación entre las puntuaciones de un test y el constructo medido afirma Martínez Arias et al. (2006). Las teorías de los test guían y condicionan el proceso de diseño, análisis y utilización del test conforme a los avances en la psicometría y estadística, permiten una utilización rigurosa y científica (Muñiz, 2018). Las más comunes son, La Teoría Clásica de los Test (TCT) es el más antiguo y conocido modelo de medición, “se remonta a finales del siglo XIX y principios del XX”. También resalta la Teoría Respuesta al Ítem (TRI) por su eficacia (Martínez Arias et al., 2006).

2.10.1 Teoría Clásica de los Test

De acuerdo con Gómez & Joan de (2009) la teoría clásica de los Test (TCT) es la teoría de los test con la que empezó la psicometría. El énfasis del modelo de la teoría clásica está puesto en lograr la exactitud de la medida, o en su defecto, la determinación precisa del error de medición. Es por esto que Charles Spearman un psicólogo inglés propone una formulación matemática para estudiar las propiedades métricas de las puntuaciones que se asignan mediante test. Sin embargo (Muñiz, 2018) asevera que la TCT desarrollara todo un conjunto de deducciones encaminadas a estimar la cuantía del error que afecta a las puntuaciones de los test. Cabe recalcar que este modelo según Spearman consiste en asumir que la puntuación que una persona obtiene en un test que se denomina puntuación empírica representada con una x , está formada por dos componentes, por un lado, la puntuación verdadera de esa persona en el test (v) y por otro un error (e). Lo anterior se puede expresar de esta manera $X = V + e$. El objetivo principal de dichas teorías es diseñar un modelo estadístico que fundamente las puntuaciones de los test y la estimación de los errores de medida que afectan dichas puntuaciones (Muñiz, 2018).

2.10.2 Teoría Respuesta al Ítem (TRI)

La TRI emerge de estudios pioneros de Thurstone en la segunda década del siglo XX (Muñiz, 2010). Sin embargo, el nacimiento de la TRI se atribuye a los trabajos de Frederick M. Lord a mitad del siglo XX. Particularmente, su trabajo de tesis doctoral *A theory of test scores* en 1953 dirigida por Gulliksen, dio el fundamento robusto a la TRI. De igual forma, Birnbaum en (Muñiz., 2018) de modelos logísticos para la evaluación psicológica. Seguidamente, en 1960 Georg Rasch publicó su libro sobre el modelo logístico de un parámetro. Más adelante en 1968 Lord y Novick publican un libro en que presentan aplicaciones prácticas a los modelos de la TRI, dando mayor fundamento a la teoría (Muñiz, 2018). Sin embargo, debido a que los modelos de la TRI implicaban complejos procesos matemáticos, fue hasta el desarrollo de software apropiado, en los años 80 (como BICAL, LOGISTIC, BILOG, MULTILOG, MICROCAT) que comenzó su aplicabilidad (Martínez Arias et al., 2006). Para finales del siglo XX, los modelos basados en TRI experimentaron una acelerada expansión y consolidación. Finalmente, se resalta el libro publicado por Frederic Lord en 1980 *Aplicaciones de la Teoría de Respuesta a los Ítems a problemas prácticos del uso de los test* en el que se sintetizan los desarrollos teóricos y modelos de la TRI. En la actualidad, la TRI se ha popularizado y ha tomado gran preponderancia en el mundo de la medición psicológica y educativa (Muñiz, 2010).

2.10.2.1 Características desde la TRI

La TRI resuelve importantes problemas en la medición psicológica que no encontraban solución en la Teoría Clásica (Muñiz, 2010). Su gran contribución y objetivo principal es obtener mediciones invariantes respecto de las personas implicadas y más allá de los ítems o instrumentos utilizados (Attorresi et al., 2009).

La razón de llamarse TRI es porque su enfoque es en el análisis del ítem individual y sus propiedades, más que en el test (Muñiz, 2010). La TRI busca establecer una relación entre el comportamiento de un sujeto al contestar un ítem y el rasgo latente o habilidad subyacente a esa conducta (Attorresi et al., 2009). Los principios básicos de esta teoría son: 1. El patrón de respuestas en un conjunto de ítems en un test es explicado por un rasgo latente, simbolizado por θ . El cual es estimado a partir del patrón de respuestas del conjunto de ítems esclarece Attorresi et al. (2009).

2. Existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos (Hidalgo-Montesinos, 2016).
3. La puntuación verdadera (el nivel de aptitud, rasgo, habilidad, competencia, etc.), es independiente del ítem y del test utilizado para estimarlo. Por lo tanto, la respuesta del sujeto depende únicamente del rasgo latente (Matas Terron, 2010).
4. Invarianza de los parámetros de los ítems respecto a la muestra que se calcula. Es decir, que los parámetros del ítem no cambian, aunque las personas que contesten sean distintas (Matas Terron, 2010).

Desde la TRI se postula que para cada nivel de rasgo existe una probabilidad asociada de contestar correctamente al ítem. Esta probabilidad es pequeña para sujetos con bajo nivel de rasgo, y alta para sujetos con altos niveles de rasgo (Matas Terron, 2010). La relación entre la respuesta de un sujeto a un ítem y el rasgo latente puede describirse como una función monótona creciente, expresada en términos probabilísticos mediante distintos modelos matemáticos (Matas Terron, 2010). Particularmente, se basa en análisis de regresión de datos multivariante que enlazan una variable respuesta con otra no observada. Normalmente, este análisis es calculado mediante funciones de probabilidad normal acumulada y logística, y es representado gráficamente mediante la Curva Característica del Ítem (CCI) afirma Atorresi et al. (2009); Matas, 2010).

El comportamiento de un individuo en un ítem es en función no solo de su nivel de aptitud, sino también de las características específicas del ítem (Martínez Arias et al, 2006). Por tanto, la forma de la CCI puede ser determinada por los siguientes parámetros del ítem:

- Localización o dificultad del ítem (parámetro b): Representa el nivel de dificultad del ítem y se define como el nivel de rasgo para el cual un individuo tiene una probabilidad del 50% de responder correctamente al ítem. Este describe dónde está situado el ítem en la escala de aptitud, por lo que se expresa en la misma escala que el rasgo latente, correspondiente al punto en el eje x de máxima pendiente de la CCI. La escala adoptada para medir θ tiene un rango teórico de $-\infty$ a $+\infty$, con media 0 y desvío típico 1. Cuanto mayor es el valor de b más difícil es el ítem (Muñiz, 2010; Navas Ara, 2001).

- Discriminación del ítem (parámetro a): Mide la potencia discriminatoria para identificar los sujetos en función del nivel en el rasgo latente que poseen. Este es proporcional a la pendiente de la recta tangente a la CCI en el punto de máxima pendiente de esta. Tiene un rango teórico entre $-\infty$ a $+\infty$, aunque normalmente es comprendida entre -2.8 y 2.8.

2.10.2.2 Supuestos desde la TRI

- **Unidimensionalidad:** Los ítems de un test están constituidos en función de un único rasgo (Muñiz, 2010). Este rasgo es suficiente para explicar los resultados de los sujetos afirma Martínez Arias et al. (2006).

El estudio de unidimensionalidad consiste en analizar la configuración interna del test, evaluando el nivel en que los ítems encajan con el constructo medido en el test (Navas Ara M. J., 2001).

Un supuesto importante desde la TRI es que los test son unidimensionales (aunque en la actualidad existen modelos multidimensionales), es decir, que un único rasgo latente es responsable explicar las respuestas de un sujeto ante un conjunto de ítems. Este supuesto es muy importante para obtener buenas mediciones y obtener evidencias de validez de estructura interna del test (Martínez Arias et al., 2006). Sin embargo, en la práctica, la unidimensional no podrá cumplirse totalmente, ya que la medición psicológica siempre involucra un conjunto de elementos interrelacionados que influyen en el comportamiento de respuesta a un conjunto de ítems (Burga Leon, 2006). No obstante, el supuesto se considera satisfecho cuando un solo factor dominante es capaz de explicar las respuestas del test y las relaciones entre ítems de acuerdo con Aune et al. (2019).

- **Independencia Local:** La respuesta a un ítem no está condicionado a la respuesta de otro ítem cuando se considera un mismo nivel en el rasgo. Es decir, que las respuestas a distintos ítems son estadísticamente independientes (Atorresi et al., 2009; Matas, 2010).

En cuanto a estos dos últimos supuestos, se argumenta que ambos están íntimamente relacionados, ya que, si se cumple la unidimensionalidad, se cumple el supuesto de independencia local (Muñiz, 2010). Esto es, si se confirma que dos ítems no son independientes significa que otro factor ajeno al que se pretende medir incide en la estimación del nivel del rasgo violando el supuesto de unidimensionalidad (Atorresi et al., 2009).

2.10.2.3 Modelos de la TRI

El análisis usando TRI proporciona una visión de la relación entre el nivel en el rasgo de un individuo y las características de los ítems. Tienen el objetivo de predecir la probabilidad de respuesta a un ítem basándose en diferentes parámetros de los ítems y el nivel de rasgo latente. La TRI proporciona una variedad de modelos que permite trabajar con test tanto unidimensionales como multidimensionales y con distintos formatos de respuesta (Hidalgo-Montesinos, 2016).

Como ya se ha mencionado, los modelos de TRI asumen que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar a estos, dicha función se expresa mediante la CCI (Muñiz, 2010). Los dos tipos funciones más utilizadas dentro del marco de la TRI son la función logística y la curva normal acumulada. La función logística tiene muchas ventajas sobre la curva normal, principalmente por su mayor tratabilidad matemática (Muñiz, 2018).

El modelo logístico puede clasificarse en función del número de parámetros: modelo de un parámetro, dos parámetros y tres parámetros (Matas Terron, 2010). Es importante señalar que los modelos logísticos asumen que la respuesta a los ítems es codificada dicotómicamente (1= acierto y 0= error), aunque involucre varias opciones de respuesta (Muñiz, 2018).

2.10.2.4 El modelo logístico de un parámetro

El modelo logístico de un parámetro o modelo de Rasch, fue formulado por George Rasch en 1960, de donde se le adjudica su nombre (Muñiz, 2018). Según este modelo, la respuesta a un ítem viene determinado solamente por el nivel del rasgo y la dificultad del ítem (parámetro b_i). En el modelo de Rasch la CCI adquiere la siguiente expresión:

$$\text{Donde: } P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

b_i = Índice de dificultad del ítem i

e = Base de los logaritmos neperianos

D = Constante (cuando $D = 1.7$, los valores de la función logística apenas difieren de los de la curva normal acumulada).

Este modelo se basa en que, conocido el índice de dificultad de un ítem b , y el nivel de rasgo del individuo θ , es posible predecir la probabilidad $P(\theta)$ de acertar el ítem. El índice de dificultad b es aquel valor de θ para el cual $P(\theta) = 0.5$, en otras palabras, es el punto de inflexión de la curva (Muñiz, 2018).

2.10.2.5 El modelo logístico de dos parámetros

El **Modelo Logístico de Dos Parámetros (ML2P)** es una de las formulaciones fundamentales dentro de la **Teoría de Respuesta al Ítem (TRI)** y se utiliza ampliamente en la construcción y análisis de pruebas educativas y psicológicas. Este modelo describe la relación entre la habilidad latente de un individuo (θ) y la probabilidad de responder correctamente a un ítem, considerando dos parámetros clave: **discriminación (a)** y **dificultad (b)** (Hambleton et al. 1991). El **parámetro de discriminación (a)** indica la capacidad del ítem para diferenciar entre individuos con distintos niveles de habilidad, siendo más efectivo cuanto mayor sea su valor. Por otro lado, el **parámetro de dificultad (b)** representa el nivel de habilidad en el cual la probabilidad de responder correctamente al ítem es del 50% (Embretson, S. E., & Reise, S. P., 2000). Matemáticamente, la función logística del ML2P se expresa como:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

donde:

- donde $P(\theta)$ es la probabilidad de una respuesta correcta y e es la base del logaritmo natural (Baker, F. B., & Kim, S. H., 2017).

El parámetro de discriminación (a)

Representa qué tan bien un ítem diferencia entre individuos con diferentes niveles de habilidad (θ). Valores altos de (a) indican que el ítem es muy sensible a las diferencias entre los examinados, es decir, que distingue eficazmente entre personas con habilidades bajas y altas.

- **Si a es alto ($a > 1.0$)**, el ítem tiene una gran capacidad de discriminación y es útil para evaluar con precisión la habilidad del examinado.
- **Si a es bajo ($a < 0.5$)**, el ítem discrimina de manera deficiente, lo que sugiere que su inclusión en la prueba podría ser cuestionable (Baker, F. B., & Kim, S. H., 2017).

En términos gráficos, un ítem con un (a) alto tendrá una **curva característica de ítem (CCI)** más inclinada, lo que indica que la transición de respuestas incorrectas a correctas ocurre en un rango más reducido de habilidad. En cambio, un ítem con un (a) bajo mostrará una pendiente menos pronunciada, reflejando una discriminación más baja.

Parámetro de Dificultad (b)

El **parámetro de dificultad (b)** representa el nivel de habilidad necesario para que la probabilidad de responder correctamente al ítem sea del 50%.

- **Si b es bajo ($b < 0$)**, significa que el ítem es fácil y que incluso examinados con baja habilidad pueden responderlo correctamente.
- **Si b es alto ($b > 0$)**, indica que el ítem es difícil y que solo los individuos con un alto nivel de habilidad podrán responderlo correctamente.

En una **curva característica de ítem (CCI)**, el valor de (b) se representa en el eje de habilidad (θ), mostrando el punto en el que la curva cruza el 50% de probabilidad de respuesta correcta.

- A diferencia del **Modelo de Un Parámetro (Rasch)**, el ML2P permite que los ítems varíen en términos de su capacidad para discriminar entre examinados, lo que lo hace más flexible en la construcción de pruebas psicométricas (Van der Linden, W. J., & Hambleton, R. K., 2013). Su aplicación es especialmente útil en pruebas adaptativas computarizadas y en la calibración de ítems en evaluaciones de gran escala, donde se requiere precisión en la estimación de la habilidad del evaluado (Lord, 1980). El modelo 2PL es ampliamente utilizado en evaluaciones educativas y psicológicas debido a su capacidad para estimar ítems con diferentes niveles de discriminación y dificultad, lo que permite construir pruebas más precisas y justas (Hambleton R. K., 1991). A diferencia del **Modelo de Un Parámetro (1PL o Rasch)**, que asume que todos los ítems tienen el mismo poder discriminativo, el **Modelo 2PL** incorpora el **parámetro de discriminación (a)**, lo que permite ajustar la influencia de cada ítem en la evaluación de la habilidad del examinado (Baker, F. B., & Kim, S. H., 2017). Esta propiedad es especialmente valiosa en pruebas de alto impacto, como el **TOEFL**, **PISA** y **pruebas adaptativas computarizadas (CAT)**, donde es crucial diferenciar con precisión a los examinados en distintos niveles de habilidad (Van der Linden, 2010). Desde una perspectiva psicométrica, el **Modelo 2PL** mejora la equidad de las evaluaciones, ya que

permite que los ítems varíen en su capacidad de discriminar entre individuos con diferentes niveles de competencia, evitando sesgos que podrían afectar la interpretación de los resultados (Embretson, S. E., & Reise, S. P., 2000). Además, facilita la construcción de **bancos de ítems calibrados**, donde cada pregunta es analizada en función de sus propiedades estadísticas antes de ser incluida en una prueba final (Lord, 1980). Esto es particularmente útil en evaluaciones longitudinales o adaptativas, ya que permite seleccionar ítems que maximicen la información obtenida de cada examinado sin necesidad de aplicar pruebas excesivamente largas o redundantes (Van der Linden, W. J., & Hambleton, R. K. , 2013). Otro beneficio clave del **Modelo 2PL** es su aplicabilidad en contextos educativos y clínicos. En educación, se emplea en la evaluación de **competencias lectoras, matemáticas y cognitivas**, permitiendo a los diseñadores de pruebas construir escalas de medición más sensibles a las diferencias individuales en el aprendizaje (Reise, S. P., & Revicki, D. A. , 2015). En el ámbito de la **psicología y la salud mental**, se ha aplicado en la construcción de pruebas de evaluación de la **ansiedad, depresión y trastornos del desarrollo**, proporcionando mediciones más ajustadas a las diferencias individuales entre los pacientes (De Ayala, R. J., 2009).

Aplicaciones del Modelo 2PL

- El modelo 2PL es utilizado en diversas áreas como:
- **Pruebas educativas estandarizadas:** Se emplea en evaluaciones como PISA, TOEFL y GRE para estimar la habilidad de los estudiantes de manera más precisa (OECD, 2019).
- **Pruebas adaptativas computarizadas (CAT):** Facilita la selección de ítems adecuados para cada evaluado en función de su nivel de habilidad (Van der Linden, 2010).
- **Validación y calibración de ítems:** Permite analizar la calidad de los ítems en términos de dificultad y discriminación, asegurando que midan lo que se pretende evaluar (Hambleton et al., 1991).

2.10.2.6 El modelo logístico de tres parámetros

El modelo de tres parámetros tiene su origen con Birnbaum entre los años de 1957 a 1968. El modelo asume que la CCI es determinada por tres parámetros: dificultades, discriminación y

pseudoazar. Este último se define como la probabilidad de acertar el ítem al azar cuando no se conoce la respuesta $P_i(\theta) = -\infty$ (Muñiz, 2018).

El modelo se expresa de la siguiente forma:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{Da_i(\theta - b_i)}}$$

Donde

$P_i(\theta)$ = Probabilidad de acertar el ítem i a determinado nivel θ

θ = Valores de la variable medida

c_i = Índice de pseudo -azar del ítem i

b_i = Índice de dificultad del ítem i

a_i = Índice de discriminación del ítem i

e = Base de los logaritmos neperianos

D = Constante ($D = 1.7$)

Evidentemente, al aumentar la probabilidad de acertar por azar, la probabilidad de respuestas correctas aumenta (Muñiz, Introducción a la Psicometría: Teoría Clásica y TRI, 2018).

Por último, la TRI permite evaluar hasta qué punto el modelo alcanzado representa bien a los datos observados. La evaluación del ajuste se basa en análisis de la discrepancia entre las CCIs teóricas, dadas por el modelo y las CCIs empíricas (Attorresi et al., 2009).

Para evaluar el ajuste de cada ítem al modelo se puede utilizar el estadístico χ^2 , que evalúa el grado de similitud entre las frecuencias de respuestas del modelo predictivo y las del modelo empírico por categoría de respuesta al ítem. Valores estadísticamente significativos indicarían que existen discrepancias entre el modelo predictivo y el observado, y por lo tanto que un ítem determinado no se ajusta al modelo (Toland, 2013).

CAPÍTULO III

3. Marco Contextual

Este marco contextual revisa y analiza diversas evaluaciones internacionales y nacionales, haciendo énfasis especial en la situación actual en Honduras, donde se identifican tanto avances significativos como desafíos persistentes en la medición y desarrollo de competencias lingüísticas en inglés. Con base en el análisis de experiencias internacionales exitosas y las limitaciones observadas en el contexto nacional, se establecen las bases para reflexionar sobre la importancia de implementar evaluaciones propias adaptadas al contexto hondureño, que favorezcan el desarrollo integral y la mejora continua del proceso educativo.

3.1 Evaluaciones Estandarizadas

Didier (2019) afirma que las pruebas estandarizadas internacionales han sido implementadas como una estrategia global para medir la calidad educativa, promoviendo la rendición de cuentas y la competencia entre países. Sin embargo, algunos estudios critican que estos exámenes tienden a priorizar resultados cuantificables sobre el desarrollo integral del estudiante, limitando la innovación pedagógica y reduciendo el aprendizaje a lo medible.

Mientras que Demarchi (2020) sostiene que las pruebas estandarizadas internacionales, son consideradas como una evaluación comparativa que realizan las agencias y organizaciones internacionales, se trata de instrumentos estandarizados de medición aplicados en varios países de manera conjunta, para recolectar información sobre los aprendizajes conceptuales y procedimentales de distintas áreas, además de indagar en los distintos factores que influyen en el aprendizaje de los estudiantes. Estas evaluaciones ayudan a identificar las habilidades, destrezas y debilidades que poseen los estudiantes en diferentes áreas de su formación educativa, al mismo tiempo ayudan a fortalecer el conocimiento que a futuro deben poseer para tener un mejor rendimiento académico. De tal manera se espera que las pruebas a gran escala como ser las pruebas nacionales e internacionales puedan ser solidas que reflejen datos válidos y así puedan aportar al beneficio de los y las estudiantes del sistema educativo. Lira & Gaete (2024) señalan que las pruebas estandarizadas internacionales en educación han sido adoptadas como instrumentos de medición de la calidad educativa en diversos países. Se argumenta que su aplicación busca ofrecer un diagnóstico comparativo del desempeño

estudiantil en competencias clave como matemáticas, lectura y ciencias, lo que permite a los gobiernos diseñar políticas educativas basadas en evidencia empírica. No obstante, también se advierte que estas pruebas pueden homogenizar la enseñanza y no siempre reflejan la diversidad de contextos educativos.

3.2 Principales Pruebas para Evaluar la Lectura en inglés en Educación Primaria

A nivel internacional, existen diversas pruebas diseñadas para medir la lectura en inglés en niños. Algunas de las más utilizadas son:

Young Learners English (YLE) – Cambridge English

El examen Young Learners English (YLE), desarrollado por Cambridge English, está diseñado para niños de educación primaria y evalúa la comprensión lectora mediante actividades lúdicas y textos sencillos. Se divide en tres niveles:

- **Starters (A1):** Evaluación de vocabulario y frases básicas.
- **Movers (A1-A2):** Lectura de textos simples con apoyo visual.
- **Flyers (A2):** Comprensión de textos narrativos cortos.

Según Cambridge Assessment (2021) esta prueba se adapta al ritmo de aprendizaje de los niños, asegurando que el contenido sea accesible y relevante para estudiantes de primer ciclo de primaria.

El TOEFL Primary, desarrollado por ETS, mide la capacidad de los niños para comprender textos en inglés. Evalúa:

- Reconocimiento de vocabulario básico en contexto.
- Comprensión de oraciones y textos cortos.
- Capacidad para identificar información clave en una lectura.

Según ETS (2020), esta prueba está alineada con el Marco Común Europeo de Referencia para las Lenguas (CEFR) y puede ser utilizada para monitorear el progreso de los estudiantes en escuelas públicas de educación primaria.

Evaluación de la Lectura según el Marco de PIRLS

Según Mullis et al (2023) el estudio PIRLS (Progress in International Reading Literacy Study) aplicado a niños de 4° grado, ha identificado estrategias clave en la evaluación de la lectura en niños, las cuales pueden adaptarse al contexto hondureño:

- Identificación de información específica en un texto corto.
- Uso de imágenes y pistas contextuales para comprender significado.
- Evaluación de la fluidez lectora mediante actividades orales y escritas.

3.3 Otras pruebas estandarizadas

- **Primary English Test (PET)** actualmente conocido como B1 Preliminary, es una prueba de inglés desarrollada por Cambridge Assessment English, diseñada para evaluar la competencia lingüística en nivel intermedio (Cambridge Assessment, 2024).
- Tajik (2025) examina la integración de la inteligencia emocional impulsada por inteligencia artificial (AI-EI) en plataformas de aprendizaje de idiomas para mejorar las habilidades orales en inglés mediante un diseño de métodos mixtos, la investigación evaluó la eficacia de un sistema innovador basado en **Amazon Alexa-Speak**, el cual proporciona retroalimentación en tiempo real adaptada al estado emocional de los estudiantes. concluye que la integración de AI-EI en plataformas de aprendizaje de idiomas representa un avance paradigmático en la enseñanza del inglés, al abordar simultáneamente los aspectos cognitivos y emocionales del proceso de adquisición del lenguaje. Tajik (2025) concluye que la integración de AI-EI en plataformas de aprendizaje de idiomas representa un avance paradigmático en la enseñanza del inglés, al abordar simultáneamente los aspectos cognitivos y emocionales del proceso de adquisición del lenguaje.
- **Trinity GESE (Graded Examinations in Spoken English):** La transformación de las prácticas evaluativas en el área de artes del lenguaje en inglés requiere un enfoque que no solo mida la competencia lingüística de los estudiantes, sino que también valore su diversidad cultural y su capacidad crítica. Según Moore y Bawany (2025), la herramienta **Try, Translate, Relate (TTR)** constituye una alternativa innovadora a los modelos tradicionales de evaluación, al permitir que los estudiantes participen activamente en la construcción de enunciados de prueba más alineados con sus realidades. Este modelo consta

de tres fases: **Try**, donde los estudiantes interpretan libremente el contenido a evaluar; **Translate**, donde reformulan enunciados de evaluación estandarizados en términos accesibles; y **Relate**, donde generan nuevas preguntas que reflejan su contexto cultural y lingüístico. A través de este proceso, los estudiantes no solo desarrollan una mayor comprensión de los criterios de evaluación, sino que también adquieren una conciencia crítica sobre los sesgos en las pruebas estandarizadas. Así, la evaluación deja de ser un mecanismo unidireccional impuesto desde estructuras de poder académico y se convierte en una herramienta de aprendizaje colaborativo y equitativo (Moore, 2025).

- **Programme for International Student Assessment (PISA)**

Aunque generalmente se aplica en secundaria, PISA también mide habilidades en inglés en niveles primarios en algunos países. El estudio de Ficayuma (2024) evidencia una discrepancia significativa entre el currículo de inglés en Indonesia y los estándares evaluativos de exámenes internacionales como **PISA** y **TOEFL**, lo que podría impactar la preparación de los estudiantes para estas pruebas. Mientras que **PISA** enfatiza textos informativos y analíticos, como descripciones, reportes y exposiciones, el currículo escolar prioriza la enseñanza de textos descriptivos y procedimentales, dejando de lado géneros argumentativos y transaccionales esenciales en la evaluación internacional. Además, se identificaron **siete competencias clave de alfabetización crítica** necesarias para la comprensión lectora en PISA, como la capacidad de distinguir hechos de opiniones y evaluar la fiabilidad de las fuentes, habilidades que no se desarrollan suficientemente en el sistema educativo indonesio. Como solución, la autora propone integrar enfoques basados en **Géneros Textuales (GBA)** y **Lingüística Sistémico-Funcional (SFL)** para fortalecer la enseñanza del inglés y preparar mejor a los estudiantes para exámenes estandarizados a nivel internacional según Ficayuma (2024).

- **Estudio Europeo de Competencia Lingüística (EECL):** Promovido por la Unión Europea, este estudio evalúa las competencias lingüísticas de los estudiantes en varios países europeos, utilizando escalas basadas en el MCER. Cada país tiene sus propias evaluaciones nacionales que pueden estar alineadas con el MCER. Según el Estudio Europeo de Competencia Lingüística (EECL), promovido por la Unión Europea, la evaluación de las competencias lingüísticas de los estudiantes en diversos países europeos se basa en escalas alineadas con el Marco Común Europeo de Referencia para las Lenguas (MCER). Ashton et al. (2012)

afirma que el EECL permite comparar el nivel de competencia en inglés entre países europeos, facilitando la movilidad académica y profesional dentro de la región.

- **EF Standard English Test (EF SET)** es una prueba de inglés en línea, estandarizada y gratuita, diseñada para evaluar las habilidades de comprensión lectora y auditiva. Aunque está orientada principalmente a adultos para certificación profesional, también puede ser útil para adolescentes que deseen medir su nivel de inglés.

El EF SET ofrece diferentes versiones de la prueba:

1. **EF SET Certificate™ (50 minutos):** Proporciona un análisis detallado de tus habilidades de comprensión lectora y auditiva, alineado con los niveles del Marco Común Europeo de Referencia (MCER) (EF SET, 2025).
 2. **EF Standard English Test: Quick Check (15 minutos):** Una prueba rápida que ofrece una estimación general de tu nivel de inglés, ideal si dispones de poco tiempo según el sitio de EF SET (2025).
- **Pearson Test of English (PTE) Young Learners:** Según Höl y Fidan (2023) el PTE Young Learners evalúa la competencia lingüística en niños de 6 a 13 años, centrándose en gramática, vocabulario, comprensión lectora y auditiva. Este examen proporciona un marco estructurado para evaluar las habilidades comunicativas en inglés y su alineación con los estándares internacionales así lo establece Höl y Fidan (2023).

3.4 Evaluaciones estandarizadas en Latinoamérica

En América Latina, diversas pruebas estandarizadas evalúan la competencia lectora en inglés como lengua extranjera.

1. **Costa Rica:** El Ministerio de Educación Pública (MEP) aplica la Prueba Nacional Estandarizada de Lenguas Extranjeras a estudiantes de último año de secundaria. Esta evaluación digital mide competencias de comprensión lectora y auditiva en inglés o francés, según corresponda. Para obtener el título de Bachiller en Educación Media, los estudiantes deben alcanzar al menos el nivel A1 del Marco Común Europeo de Referencia para las Lenguas (MEP, 2024).

Chile: la Prueba de Acceso a la Educación Superior (PAES) incluye una evaluación de Competencia Lectora en inglés, aunque centrada en la lengua materna, como parte del proceso de admisión universitaria (Pais, 2024).

Colombia: desde el año 2006, se incluyó el inglés como un componente fundamental en la prueba de Estado, elaborada y administrada por el Instituto Colombiano para la Evaluación de la Educación o ICFES, mejor conocida como la Prueba Saber 11. Esta prueba evalúa los conocimientos de los estudiantes de grado 11, próximos a culminar la educación secundaria, en diferentes áreas de conocimiento según Gutiérrez et al, (2021) “El propósito principal de las pruebas SABER es contribuir al mejoramiento de la calidad de la educación colombiana mediante la realización de evaluaciones aplicadas periódicamente para monitorear el desarrollo de las competencias básicas en los estudiantes de educación básica, como seguimiento de calidad del sistema educativo (Cortes Velandia, 2018).

México: Examen de Certificación de Lengua Extranjera: Aunque no se menciona una prueba específica en las fuentes consultadas, es común que las universidades mexicanas requieran certificaciones de inglés, como el TOEFL o el PTE, para procesos de admisión o graduación (UPAEP, 2025). Además, el **Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)** requiere que los aspirantes a programas de posgrado presenten una puntuación mínima en exámenes estandarizados de inglés, como el TOEFL, como parte de sus criterios de admisión (UPAEP, 2025).

3.2.5 Evaluaciones en la educación básica de Honduras

En el sistema educativo de Honduras, los estudiantes de **educación básica** son evaluados mediante diversas pruebas diseñadas para medir su desempeño en áreas fundamentales del conocimiento. Estas evaluaciones incluyen tanto pruebas nacionales como internacionales, orientadas a monitorear el progreso académico y a identificar áreas de mejora en el aprendizaje.

3.2.5.1 Evaluaciones Nacionales

Las pruebas diagnósticas constituyen una de las estrategias más utilizadas a nivel nacional para evaluar el rendimiento estudiantil en asignaturas clave como Español, Matemáticas e Inglés. Su objetivo es identificar el nivel de competencia de los estudiantes y proporcionar información para el diseño de intervenciones pedagógicas adecuadas (Secretaría de Educación de Honduras, 2021). Sin embargo, el monitoreo del aprendizaje del inglés ha sido limitado, lo que genera un vacío en la medición del dominio de este idioma en el contexto educativo hondureño (Reyes Vásquez, 2023).

3.2.5.2 Evaluaciones Internacionales

Honduras ha participado en diversas pruebas internacionales que permiten establecer comparaciones con el desempeño de otros países. Entre ellas, el **Estudio Regional Comparativo y Explicativo (ERCE)**, coordinado por la **UNESCO**, es una de las evaluaciones más relevantes en la región. Este estudio mide el rendimiento de los estudiantes de **tercero y sexto grado** en áreas como **Lectura, Escritura, Matemáticas y Ciencias Naturales**. Los resultados obtenidos en el **ERCE 2019** reflejan que Honduras se encuentra por debajo del promedio regional en comprensión lectora, evidenciando desafíos significativos en el desarrollo de esta habilidad (UNESCO, 2021).

3.2.5.3 Evaluaciones en el Área de Inglés

Si bien existen evaluaciones nacionales e internacionales que miden el desempeño en áreas fundamentales, la evaluación de la competencia en inglés en educación básica en Honduras sigue siendo limitada. La ausencia de pruebas estandarizadas específicas para el inglés representa un desafío en la implementación de estrategias pedagógicas efectivas que permitan fortalecer el aprendizaje de este idioma. No obstante, Honduras ha mostrado avances significativos en el dominio del inglés en los últimos años. Según el Índice EF de Dominio del Inglés (EF EPI) 2023, el país se posicionó como el de mejor desempeño en Centroamérica y el segundo en Latinoamérica, con una puntuación de 545, superando la media global de 493 (EF Education First, 2023). Estos resultados reflejan un progreso en la enseñanza y aprendizaje del idioma inglés, aunque persiste la necesidad de desarrollar evaluaciones estandarizadas propias que permitan medir con mayor precisión las competencias lingüísticas de los estudiantes hondureños y su alineación con los estándares internacionales.

En este contexto, investigaciones recientes realizadas en Honduras han identificado retos específicos relacionados con la comprensión lectora en inglés como lengua extranjera. Un estudio de caso llevado a cabo en estudiantes del III Ciclo de Educación Básica en La Esperanza, Intibucá, evidenció que gran parte de los estudiantes presentan deficiencias importantes en habilidades clave como la extracción e interpretación de información y la identificación de ideas centrales y secundarias en textos en inglés (Rodríguez et al., 2023).

De esta forma, es imprescindible avanzar hacia el desarrollo de instrumentos de evaluación que integren situaciones auténticas, relacionadas con el contexto cultural y social de

Honduras. Este enfoque podría ayudar a evaluar más efectivamente la competencia lectora real y pragmática de los estudiantes, ofreciendo datos más precisos que permitan orientar mejor las políticas educativas y prácticas pedagógicas hacia una mejora continua en el aprendizaje del inglés como lengua extranjera.

DEGT-UNAH

CAPITULO IV

4. Metodología

En esta sección se detallan los procedimientos metodológicos realizados en el estudio de construcción de la prueba de inglés como lengua extranjera (EFL) que evalúa la competencia lectora en estudiantes de primer ciclo de educación primaria. Este apartado incluye el tipo de investigación, caracterización de la muestra, operacionalización de variables, descripción del instrumento, además del procedimiento y el plan de análisis que se ejecutó en el proceso del estudio. En el contexto hondureño, la construcción de una prueba de competencia lectora en inglés como lengua extranjera para estudiantes de 4to grado requiere un proceso de validación exhaustivo, debido a las diferencias en el acceso al aprendizaje del idioma y los niveles de exposición al inglés en distintas regiones del país. Por ello, este estudio sigue el modelo de Muñiz y Fonseca-Pedrero (2019) el cual establece un procedimiento de pasos que permite garantizar la calidad psicométrica de los instrumentos de evaluación.

4.1 Tipo de investigación

Esta investigación consiste en un diseño instrumental, ya que se realiza un análisis psicométrico para reunir evidencias de validez del test con un enfoque cuantitativo, mismo que nos permite ser lo más objetivos posibles sin mostrar algún grado de subjetividad a la hora de interpretar los resultados, es de tipo exploratorio, ya que como lo menciona Hernandez-Sampieri et al. (2014) “Los estudios exploratorios se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes”.

4.2 Población y Muestra

La población del estudio está conformada por estudiantes de cuarto grado de educación primaria de instituciones educativas del Distrito Central de Francisco Morazán. Para la selección de la muestra, se consideraron dos centros educativos públicos pertenecientes al Distrito Escolar No. 11: la Escuela 4 de Junio, ubicada en la Colonia Esperanza, y el Centro de Educación Básica Juan Ramón Molina, localizado en la Colonia San Miguel.

El procedimiento de muestreo utilizado fue por conveniencia, debido a la accesibilidad y disponibilidad de los participantes dentro del contexto educativo seleccionado. La muestra final estuvo conformada por 300 estudiantes, distribuidos de la siguiente manera:

Escuela	Niños	Niñas
4 de Junio de la colonia La Esperanza	51	49
Centro de Educación Básica Juan Ramon Molina	102	98

Dado que el estudio busca evaluar la competencia lectora en inglés en un contexto donde la enseñanza de este idioma es limitada, se establecieron los siguientes criterios de inclusión para garantizar la homogeneidad de la muestra:

- Estudiantes matriculados en cuarto grado de educación primaria en las escuelas seleccionadas.
- Estudiantes cuya lengua materna sea el español y que no hayan recibido instrucción formal en inglés en niveles previos.
- Estudiantes que hayan cursado su educación preescolar y primaria en español en instituciones públicas.
- Estudiantes de nacionalidad hondureña con padres hondureños, cuya lengua nativa sea exclusivamente el español.

4.3 Variables e indicadores

La variable de interés en este estudio es la competencia lectora en inglés como lengua extranjera. Desde una perspectiva psicopedagógica, la competencia lectora se define como la capacidad de un individuo para decodificar, comprender, interpretar y evaluar textos escritos, aplicando estrategias cognitivas y metacognitivas que permitan extraer significado del material leído (Grabe, W., & Stoller, F. L., 2002).

4.3.1 Matriz de Operacionalización de Variables

Variable	Definición Conceptual	Definición operacional	Dimensiones	Indicadores
Competencia Lectora del inglés	Habilidad para comprender y procesar Información de textos escritos en inglés	Puntajes obtenidos de la prueba de competencia lectora del inglés (EFL) en niños de primer ciclo de educación primaria	Dimensión de competencia lectora de inglés como lengua extranjera (EFL)	32 ítems de selección múltiple 11 ítems de respuesta abierta 1 ítem de secuencia de eventos 2 ítems de selección de elementos 4 ítems de Completación

Nota: Elaboración propia

4.4 Instrumento de medición

En este estudio, se diseñó un total de 50 ítems, estructurados con base en los estándares de inglés del Currículo Nacional de Honduras, asegurando su alineación con los objetivos de aprendizaje y competencias esperadas en la enseñanza del inglés como lengua extranjera en 4to grado de educación primaria. El instrumento aplicado en el marco de este estudio fue diseñado sistemáticamente por el docente investigador en estricto apego a los lineamientos curriculares y evaluativos dictaminados por los Estándares Educativos Nacionales-Ingles 1ro. a 9no. Grado y que son propiedad de la Secretaría de Educación de Honduras (LPI-SEDUC-001-2015, 2015).

Para garantizar la coherencia entre el contenido del test y los estándares educativos, se elaboró una tabla de especificaciones, detallando la distribución de los ítems según los contenidos específicos del currículo de inglés para niños de 4to grado asegurando que la prueba midiera las competencias lectoras descritas en los lineamientos oficiales del Ministerio de Educación (LPI-SEDUC-001-2015, 2015). El test será aplicado en **formato papel y lápiz**, utilizando cuadernillos de evaluación diseñados específicamente para estudiantes de 4to grado. La decisión de utilizar un formato impreso responde a la necesidad

de garantizar accesibilidad universal. Las instrucciones para los estudiantes y docentes estarán claramente impresas en los cuadernillos de evaluación, asegurando claridad y comprensión del procedimiento de aplicación.

El diseño de los ítems se basará en principios psicométricos para asegurar la validez del test y su capacidad de medir la competencia lectora en inglés como lengua extranjera.

Tipo de ítems:

- 32 ítems de opción múltiple (3 alternativas): Evaluará competencia lectora asociada con textos cortos y con imágenes.
- 4 ítems de completación (3 alternativas): Selección de palabras completando el espacio con la opción correcta y apoyándose de la imagen o texto.
- 2 ítems de selección de elemento (3 alternativas): Selección de la respuesta correcta y marcación con una X la respuesta correcta.
- 1 ítem de secuencia: Escritura de números del 1-4 de acuerdo a la secuencia de eventos.
- 11 ítems de respuesta abierta: Escritura de la respuesta correcta.

El Formato de respuestas será el siguiente:

- Palabras cerradas con alternativas de respuesta única.
- Espacio en blanco con la palabra adecuada.
- Cuadros con espacio en blanco para marcar correctamente la respuesta correcta con un símbolo solicitado.
- Escritura de números de acuerdo a la secuencia de una serie de eventos distribuidos en el párrafo
- Escritura de palabras o frase solicitada

4.5 Plan de análisis de datos

La construcción de instrumentos de evaluación psicométrica requiere un proceso sistemático y metodológicamente riguroso para garantizar la validez, confiabilidad y equidad en la medición de las habilidades y competencias objeto de estudio.

En este estudio, la validez será evaluada a través del juicio de expertos (CVC de Hernández-Nieto) y el Análisis Factorial Exploratorio (AFE). La confiabilidad se analizará con el Alfa

de Cronbach y el Omega de McDonald, mientras que los parámetros de los ítems serán examinados mediante la Teoría de Respuesta al Ítem (TRI).

1. **Validación por Jueces Expertos, Coeficiente de Validez de Contenido (CVC)**

Siguiendo la metodología de Hernández-Nieto (2021), se empleará el Coeficiente de Validez de Contenido (CVC) como medida estadística para evaluar la pertinencia y calidad de los ítems diseñados. La interpretación del CVC, se puede realizar con la siguiente escala de valores: a) Menor a 0.6 validez y concordancia inaceptables. b) Igual o mayor de 0.6 y menor a 0.7, validez y concordancia deficientes. c) Mayor que 0.71 y menor o igual que 0.8, validez y concordancia aceptables. d) Mayor que 0.8 y menor o igual a 0.9, validez y concordancia buenas. e) Mayor que 0.9, validez y concordancia excelentes.

En el presente estudio se utilizarán cuatro expertos del área, sin embargo, en este método propuesto por Hernández-Nieto requiere que los ítems sean evaluados por al menos tres expertos en el área de estudio a los cuales se les entregaran plantillas en las que analizan cada reactivo en función de los criterios de pertinencia, claridad conceptual, redacción y terminología, respuesta correcta, distractores apropiados, niveles de dificultad y formato. Cada uno de estos aspectos será calificado por los jueces utilizando una escala tipo Likert de cinco puntos, en la que 1 representa una evaluación inaceptable y 5 una evaluación excelente. El grado de acuerdo entre los expertos se calculará mediante el **Coeficiente de Validez de Contenido (CVC)** propuesto por Hernández-Nieto (2002), el cual permite cuantificar la relevancia de los ítems en una escala de 0 a 1. Se considerará aceptable un valor de **CVC $\geq$ 0.80** para cada ítem (Hernandez-Nieto, 2002).

Luego de haber obtenido evidencia de validez de contenido mediante la evaluación de jueces expertos y haber realizado los ajustes necesarios en función de sus sugerencias, se procede ahora con la aplicación de una prueba piloto.

Prueba Piloto

Para evaluar el funcionamiento de los ítems en condiciones reales de aplicación, se llevará a cabo una prueba piloto con una muestra representativa de 300 estudiantes de 4to grado. En el contexto de esta investigación, el estudio piloto tendrá como finalidad examinar la adecuación de los ítems diseñados para medir la competencia lectora en niños de cuarto

grado de educación primaria. A través de esta prueba preliminar, se busca determinar si los reactivos son comprensibles, si las instrucciones son claras.

El test será administrado de manera grupal e individual, dependiendo de la estructura y necesidades de la institución educativa. Se procurará que todos los estudiantes completen la prueba en condiciones homogéneas, evitando interrupciones y garantizando que reciban las mismas indicaciones antes de iniciar la evaluación.

Procedimiento para la aplicación de la prueba piloto

Se aplicará el test en un entorno escolar controlado tomando en cuenta lo siguiente:

- 1. Logísticos:** Para la recolección de los datos, se seleccionará una muestra de 300 estudiantes de dos escuelas públicas del distrito #11 del Distrito Central que tuvieran características similares a la población objeto de estudio.
- 2. Éticos:** Se visitará el centro para pedir autorización, y luego de ser confirmada la participación voluntaria, los directivos firmarán una autorización de aplicación del instrumento.
- 3. Metodológicos:** La prueba de inglés se aplicará a todas las secciones de cuarto grado ya que es un instrumento que establece que debe ser aplicada a estudiantes que hayan cursado el primer nivel de primaria logrando una muestra de 300 estudiantes. Al finalizar la evaluación, los cuadernillos de prueba serán recopilados y revisados para verificar la completitud de las respuestas. Posteriormente, los datos obtenidos serán codificados y almacenados en una base de datos digital de Excel, resguardando la confidencialidad de la información.

Se implementarán medidas de seguridad en el almacenamiento de los datos, asegurando que solo el investigador tenga acceso a los resultados con el fin de garantizar el uso ético y adecuado de la información obtenida.

2. Validez estructural del test (Análisis Factorial Exploratorio - AFE)

La validez de constructo se evaluará mediante un Análisis Factorial Exploratorio (AFE), técnica que permite identificar la estructura subyacente del test y verificar si los ítems se agrupan de manera coherente en dimensiones específicas (Muñiz, 2018).

Para determinar la adecuación de los datos al análisis factorial, se emplearán los siguientes criterios:

- **Índice de adecuación muestral de Kaiser-Meyer-Olkin (KMO):** Se considerará aceptable un valor ≥ 0.70 (Kaiser, 1974).
- **Prueba de esfericidad de Bartlett:** Se verificará si los datos presentan correlaciones adecuadas para el análisis factorial (Bartlett, 1954).
- El método de extracción será Eje de Componentes Principales con rotación Promax, considerando la posible correlación entre los factores (Brown, 2015).

3. Confiabilidad del Test (Alfa de Cronbach y Omega de McDonald)

La confiabilidad del instrumento se evaluará mediante coeficientes de consistencia interna. Se utilizarán dos indicadores complementarios: el coeficiente **Alfa de Cronbach** y el coeficiente **Omega de McDonald** (Ventura-León & Caycho-Rodríguez, 2017; Muñiz & Fonseca-Pedrero, 2019).

- **Alfa de Cronbach (1951):** Este coeficiente estima la confiabilidad a partir de la correlación entre los ítems. Un valor superior a 0.70 será considerado aceptable.
- **Omega de McDonald (1999):** Se utilizará como alternativa más robusta al Alfa de Cronbach, ya que este coeficiente es más adecuado cuando los ítems presentan cargas factoriales heterogéneas (McDonald, 1999). Se considerará óptimo un valor ≥ 0.70 , aunque en algunas circunstancias pueden aceptarse valores superiores a .65 (Ventura-León, 2017).

Ambos coeficientes permitirán determinar la fiabilidad del test tanto en su totalidad como en cada dimensión identificada en el AFE (Muñiz, 2018).

4. Análisis de los Ítems mediante Teoría de Respuesta al Ítem (TRI)

El análisis de los ítems se realizará con modelos de **Teoría de Respuesta al Ítem (TRI)**, los cuales permitirán evaluar la calidad de los reactivos mediante parámetros matemáticos de dificultad y discriminación (Susan E. Embretson, 2000). Se aplicará un modelo de **dos parámetros (2PL)**, que considera:

- **Parámetro de Discriminación (a):** Indica qué tan bien un ítem diferencia entre estudiantes con alto y bajo nivel de competencia lectora. Se considerarán adecuados valores de $a \geq 0.80$ (Baker, 2001).

- **Parámetro de Dificultad (b):** Representa el nivel de habilidad necesario para responder correctamente un ítem. Se analizará si los valores de dificultad están distribuidos de manera adecuada a lo largo de la escala de competencia lectora de inglés en lengua extranjera (Hambleton R. K., 1991).

Para los análisis se utilizarán los programas Excel, JASP, Jamovi y R, RStudio.

DEGT-UNAH

CAPÍTULO V

5. Resultados de la Investigación

Este capítulo presenta los principales hallazgos derivados del análisis de los datos recolectados, en coherencia con los objetivos planteados en el estudio. En primera instancia, se describen las características estadísticas básicas de los ítems que conforman el instrumento, a través del análisis descriptivo de tendencia central y dispersión. Posteriormente, se abordan los análisis inferenciales, incluyendo la validez de constructo mediante análisis factorial exploratorio (AFE), así como la confiabilidad interna del instrumento a través del coeficiente alfa de Cronbach y el coeficiente omega de McDonald. Finalmente, se presentan los resultados derivados de la Teoría de Respuesta al Ítem (TRI), con énfasis en los parámetros de discriminación y dificultad de los ítems. En conjunto, estos análisis permiten evaluar la calidad psicométrica del instrumento en términos de validez, consistencia interna y funcionamiento de los ítems.

Antes de proceder con los análisis estadísticos, se realizó la aplicación piloto del instrumento conformado por 39 ítems, previamente validado por juicio de expertos. Esta fase tuvo como propósito evaluar el comportamiento de los ítems en condiciones reales de aplicación. Debido a limitaciones logísticas y de tiempo, el único permiso formal gestionado fue la autorización de los directores de los centros educativos participantes, quienes accedieron a facilitar el proceso de aplicación. Las constancias de autorización correspondientes se incluyen en los anexos de este documento.

5.1 Validez de contenido por jueces expertos

El primer objetivo de esta investigación consistió en analizar la validez de contenido del instrumento diseñado para evaluar la competencia lectora en inglés como lengua extranjera (EFL). Para ello, se contó con la participación de cuatro jueces expertos, quienes evaluaron cada uno de los 50 ítems inicialmente propuestos mediante el Coeficiente de Validez de Contenido (CVC) descrito por Hernández-Nieto (2002). Se estableció como criterio mínimo para aceptar los reactivos un valor de $CVC \geq 0.90$, considerado indicador de una validez excelente.

Los ítems analizados eran de opción múltiple, con tres alternativas de respuesta cada uno: una opción correcta (clave) y dos distractores, lo cual implicó evaluar un total de 50 claves y 100 distractores. Tras la evaluación de los jueces expertos, se identificó que 11 reactivos (22%) presentaron valores inferiores al criterio mínimo establecido ($CVC \leq 0.80$), lo que indicó deficiencias en aspectos específicos como claridad en la redacción, calidad de los distractores o relevancia para evaluar el constructo. Por lo tanto, estos reactivos fueron eliminados del instrumento final. En contraste, los 39 ítems restantes (78%) presentaron valores de CVC superiores al criterio mínimo (≥ 0.90), indicando una alta concordancia y validez excelente según los estándares establecidos (Hernández-Nieto, 2002; Sánchez, 2021).

Tabla 1: Resumen del análisis de validez de contenido mediante juicio de expertos (CVC)

Número total de Ítems	Ítems aceptados (CVC ≥ 0.90)	% Ítems aceptados	Ítems rechazados (CVC < 0.90)	% Ítems rechazados	CVC promedio
50	39	78%	11	22%	0.94

Nota: Elaboracion propia

5.1.2 Descriptivos de los ítems del instrumento

El análisis de los estadísticos descriptivos permitió observar una amplia variabilidad en el comportamiento de los ítems del instrumento. Algunos ítems presentaron medias particularmente altas, como *It_10* ($M = 0.87$), *It_30* ($M = 0.96$) y *It_32* ($M = 0.89$), lo que sugiere que fueron resueltos correctamente por la mayoría de los estudiantes, posiblemente indicando un bajo nivel de dificultad. Por otro lado, ítems como *It_19* ($M = 0.29$) e *It_20* ($M = 0.27$) mostraron medias bajas, lo cual podría evidenciar una mayor complejidad o poca claridad en su formulación.

Además, el ítem *It_30* presentó una desviación estándar de solo 0.19, lo que refleja una baja variabilidad en las respuestas. Este tipo de comportamiento puede limitar la capacidad del ítem para distinguir entre distintos niveles de habilidad lectora, reduciendo su valor informativo dentro del instrumento.

5.1.3 Validez Estructural (Análisis Factorial Exploratorio (AFE))

Con el objetivo de analizar la validez de constructo del instrumento, se realizó un análisis factorial exploratorio (AFE) utilizando el método de extracción de componentes principales (PA), con rotación oblicua tipo Promax y un umbral mínimo de carga factorial de .30. El análisis se aplicó a los datos recolectados de la muestra total de estudiantes que respondieron la prueba de competencia lectora en inglés como lengua extranjera (EFL).

Previo a la ejecución del AFE, se verificaron los supuestos de adecuación muestral. El índice de Kaiser-Meyer-Olkin (KMO) fue de .62, lo cual, si bien se encuentra por debajo del valor óptimo de .70, se considera aceptable en contextos exploratorios (Kaiser, 1974). La prueba de esfericidad de Bartlett fue estadísticamente significativa ($p < .001$), lo que indica que la matriz de correlaciones es adecuada para proceder con el análisis factorial (Bartlett, 1954).

Tabla 2: Modelos factoriales exploratorios en función de cargas factoriales y presencia de cargas negativas

Modelo	Nº de factores	Ítems con cargas factoriales más altas	Ítems con cargas negativas
1 factor	1	It_4 (.85), It_6 (.80), It_8 (.56), It_12 (.48), It_15 (.46)	It_19 (-.56), It_17 (-.42)
2 factores	2	F1: It_4 (.85), It_6 (.79), It_8 (.60), It_15 (.46) F2: It_18 (.84), It_16 (.76), It_12 (.55)	It_7, It_5, It_25
3 factores	3	PA1: It_4 (.84), It_6 (.76), It_8 (.62), It_15 (.48) PA2: It_18 (.86), It_16 (.74), It_12 (.59) PA3: It_3 (.60), It_20 (.59), It_9 (.45)	It_19 (-.62), It_7 (-.67), It_5 (-.51)

Nota. Se presentan los ítems con mayores cargas factoriales en cada solución, así como aquellos que mostraron cargas negativas. Las cargas negativas pueden reflejar inconsistencias en la redacción o bajo ajuste al modelo. Fuente: Elaboración Propia

5.1.4 Análisis de confiabilidad

La confiabilidad del instrumento fue evaluada mediante el coeficiente Alfa de Cronbach y el coeficiente Omega de McDonald, con el objetivo de estimar la consistencia interna de los ítems. El análisis arrojó un coeficiente **Alfa de Cronbach de 0.30**, valor considerado bajo, lo que indica una limitada consistencia entre los ítems del instrumento. Por su parte, el coeficiente **Omega de McDonald fue de 0.67**, valor cercano al umbral aceptable para estudios en fase exploratoria.

5.1.5 Estimación de parámetros psicométricos desde la TRI y su función de información (2PL)

En el marco de la Teoría de Respuesta al Ítem, se aplicó un modelo logístico de dos parámetros (2PL) con el objetivo de estimar los valores de discriminación (α) y dificultad (β) para los 39 ítems que componen el instrumento de evaluación de competencia lectora en inglés como lengua extranjera (EFL). Los resultados completos de los parámetros estimados para cada ítem se presentan en una tabla incluida en los anexos.

Los valores del parámetro de discriminación (α) oscilaron entre -2.200 y 63.434. Sin embargo, varios ítems mostraron valores atípicos, tanto negativos como extremadamente elevados, lo cual sugiere posibles problemas en la capacidad discriminativa de esos reactivos. En particular, los ítems **It_7** ($\alpha = -2.200$), **It_18** ($\alpha = 63.434$), **It_23** ($\alpha = -0.012$) e **It_24** ($\alpha = 0.006$) reflejan inconsistencias en su comportamiento, siendo potencialmente ineficaces para distinguir entre estudiantes con diferentes niveles de competencia lectora. Por otro lado, se observaron ítems con niveles adecuados de discriminación, como **It_16** ($\alpha = 3.440$) e **It_12** ($\alpha = 2.445$), lo que indica que estos reactivos presentan una buena sensibilidad para diferenciar entre los niveles de habilidad.

Respecto al parámetro de dificultad (β), los valores oscilaron entre -230.938 y 139.498, lo que evidencia una alta variabilidad en la ubicación de los ítems a lo largo del continuo de habilidad. Algunos ítems presentan valores extremadamente bajos o altos (e.g., **It_24** con $\beta = -230.938$ e **It_23** con $\beta = 139.498$), lo cual podría señalar problemas de ajuste o errores en

la estimación de estos parámetros. Los ítems con valores de dificultad moderada, como **It_6** ($b = -1.205$) e **It_11** ($b = 0.546$), podrían considerarse mejor calibrados para evaluar una gama intermedia de habilidad.

En conjunto, estos resultados permiten identificar tanto reactivos funcionales como reactivos que requieren una revisión detallada para garantizar la calidad psicométrica del instrumento.

En la Tabla 3 se presentan los cinco ítems con mayor capacidad discriminativa, lo cual sugiere que estos reactivos son particularmente eficaces para diferenciar entre estudiantes con distintos niveles de habilidad lectora en inglés. Específicamente, los ítems **It_16** ($\alpha = 3.440$) e **It_12** ($\alpha = 2.445$) evidencian un comportamiento psicométrico robusto dentro del modelo 2PL. Por otro lado, se incluyen también ítems con valores atípicos o potencialmente problemáticos, como **It_18** ($\alpha = 63.434$), **It_23** ($b = 139.498$) e **It_24** ($b = -230.938$), cuyos resultados podrían estar indicando una mala calibración, errores de codificación, o una inadecuada representación del nivel de habilidad medido.

Tabla 3: Ítems con parámetros representativos y problemáticos del modelo 2PL, estimados para el instrumento de evaluación de competencia lectora en inglés (EFL).

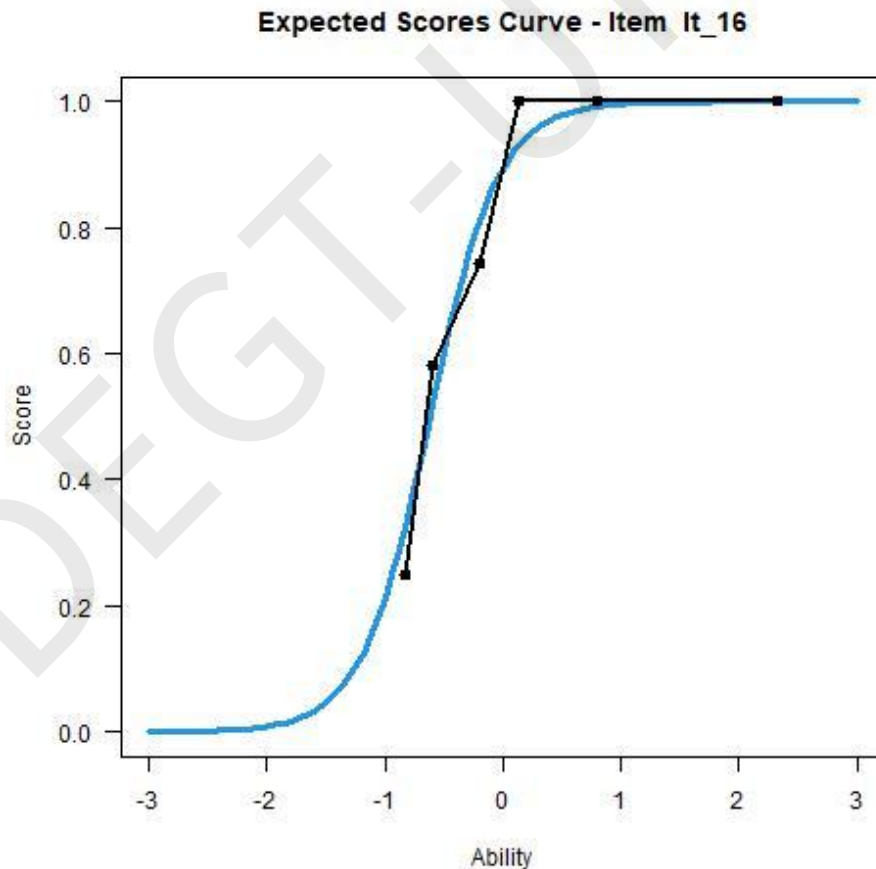
Ítem	Discriminación (α)	Dificultad (b)
It_6	0.780	-1.205
It_7	-2.200	0.350
It_12	2.445	-0.190
It_16	3.440	-0.619
It_18	63.434	-0.084
It_22	1.572	-1.582
It_23	-0.012	139.498
It_24	0.006	-230.938
It_28	-0.015	65.272
It_32	0.012	-182.022
It_36	0.034	-36.508

Nota. Elaboración propia a partir de los resultados obtenidos mediante el análisis TRI (modelo 2PL).

Figura 1

Curva característica del ítem (CCI) – It_16 estimada mediante el modelo 2PL.

Se presenta a continuación la curva característica del ítem (CCI) correspondiente al ítem **It_16**, uno de los reactivos con mejor desempeño psicométrico en el análisis TRI (modelo 2PL). Esta curva muestra la probabilidad esperada de respuesta correcta en función del nivel de habilidad (θ) del estudiante. La pendiente pronunciada observada alrededor del eje central indica una alta capacidad discriminativa ($\alpha = 3.440$), lo que implica que el ítem distingue con precisión entre estudiantes con niveles bajos y altos de competencia lectora. La forma de la curva y su ubicación reflejan que el ítem es particularmente sensible para evaluar habilidades en el rango medio-alto del continuo, lo cual lo convierte en un elemento útil para propósitos diagnósticos dentro del instrumento.



CAPÍTULO VI

6. Discusión de los Resultados

En este capítulo se discuten e interpretan los hallazgos obtenidos en la presente investigación sobre la construcción y validación de un instrumento diseñado para evaluar la competencia lectora en inglés (EFL) en estudiantes hondureños de cuarto grado. Se analizan los resultados más relevantes en relación con los objetivos planteados inicialmente, contrastándolos con estudios previos, examinando sus implicaciones educativas y psicométricas.

6.1 Hallazgos Principales (validez por Jueces Expertos)

Los resultados obtenidos en este estudio evidencian una alta validez de contenido del instrumento, con un CVC promedio de 0.94 y 39 ítems aceptados de los 50 inicialmente propuestos, lo cual indica una excelente concordancia entre los jueces expertos respecto a la pertinencia y claridad de los ítems. Estos hallazgos coinciden con lo planteado por Muñiz y Fonseca-Pedrero (2019), quienes destacan que el juicio de expertos es un procedimiento fundamental en las primeras etapas de validación de pruebas, al garantizar que los ítems reflejen adecuadamente el dominio del contenido. De forma similar, Hernández-Nieto (2002) establece que un coeficiente de validez de contenido (CVC) igual o superior a 0.90 indica una concordancia excelente, como la observada en el presente estudio. Además, estudios como el de Bazaga y Marín (2003), quienes desarrollaron la prueba TAELIS, también reportaron el uso de juicio de expertos y análisis psicométricos rigurosos, confirmando la utilidad de este procedimiento para garantizar la validez de contenido en instrumentos de evaluación lingüística. En consecuencia, los resultados obtenidos refuerzan el valor del juicio experto como un método confiable y ampliamente aceptado para asegurar que los ítems desarrollados sean representativos, comprensibles y adecuados al contexto educativo hondureño.

6.1.2 Validez Estructural (Análisis Factorial Exploratorio (AFE))

El análisis factorial exploratorio realizado con el objetivo de identificar la estructura subyacente del instrumento no arrojó una solución clara ni consistente. Se probaron modelos de uno, dos y tres factores, utilizando el método de extracción por ejes principales con rotación oblicua (Promax), considerando un umbral mínimo de .30 para la interpretación de

las cargas factoriales. Sin embargo, los resultados obtenidos evidenciaron cargas factoriales negativas, cruzadas y débiles, lo cual dificultó la interpretación de los factores y cuestionó la coherencia interna del instrumento.

El modelo de un solo factor mostró cargas elevadas en algunos ítems (por ejemplo, $It_4 = .85$; $It_6 = .80$), pero también presentó valores negativos ($It_{19} = -.56$; $It_{17} = -.42$), lo cual sugiere que los ítems no miden una única dimensión de manera homogénea. En los modelos de dos y tres factores, si bien se logró una distribución más amplia de los ítems, persistieron problemas de ambigüedad estructural, como cargas cruzadas y factores con un número reducido de ítems con cargas significativas, lo cual compromete la interpretación teórica y la estabilidad factorial.

Estos hallazgos coinciden con lo señalado por autores como Viladrich, Angulo-Brunet y Doval (2017), quienes advierten que una estructura factorial débil puede deberse a la formulación de los ítems, a una conceptualización poco clara del constructo o a la presencia de múltiples dimensiones no controladas. Asimismo, una estructura inestable podría estar vinculada al tamaño muestral o al nivel de comprensión de los ítems por parte del grupo participante. En este caso, los resultados del AFE sugieren que la versión actual del instrumento no presenta una estructura factorial sólida ni interpretable, lo cual afecta su validez de constructo en esta etapa inicial de desarrollo.

6.1.3 Confiabilidad (Consistencia interna)

En el presente estudio, el coeficiente alfa de Cronbach obtenido fue de .30, valor que se considera insatisfactorio para evaluar la consistencia interna del instrumento. Este resultado indica que los ítems incluidos no presentan una correlación suficiente entre sí como para conformar una escala confiable, lo que limita la interpretabilidad de los puntajes. Tal como planteó Cronbach (1951), este coeficiente permite estimar la homogeneidad de los ítems, y se espera que los valores sean superiores a .70 para alcanzar un nivel mínimo aceptable de fiabilidad en contextos exploratorios.

Este hallazgo es consistente con lo señalado por Oviedo y Campo-Arias (2005), quienes advierten que valores por debajo de .50 reflejan una baja cohesión entre los ítems, posiblemente atribuible a problemas de redacción, ambigüedad semántica o una estructura no unidimensional del constructo. Además, el coeficiente omega de McDonald, calculado

como medida complementaria, arrojó un valor de .67, que, aunque superior al alfa, tampoco alcanza un nivel plenamente satisfactorio.

De acuerdo con Viladrich, Angulo-Brunet y Doval (2017), el coeficiente omega resulta más apropiado en escalas con cargas desiguales o estructuras multidimensionales. No obstante, en este caso, su valor moderado refuerza la evidencia de que el instrumento no alcanza niveles adecuados de fiabilidad interna bajo los criterios psicométricos establecidos.

6.1.4 Parámetros psicométricos desde la TRI y su función de información (2PL)

El análisis psicométrico mediante el modelo logístico de dos parámetros (2PL) permitió evaluar con mayor precisión el comportamiento de los ítems que integran el instrumento de evaluación de la competencia lectora en inglés como lengua extranjera (EFL). Este modelo permitió estimar de manera independiente los parámetros de discriminación (α) y dificultad (β) de cada reactivo, proporcionando información clave sobre su adecuación para evaluar el rasgo latente propuesto.

Los resultados revelaron una alta variabilidad en los valores de α y β , lo que es consistente con instrumentos en etapa de construcción. Algunos ítems mostraron valores positivos moderados de discriminación (e.g., **It_12**, **It_16**, **It_22**), lo cual indica que tienen una adecuada capacidad para distinguir entre estudiantes con diferentes niveles de competencia lectora. Este hallazgo respalda la utilidad de estos ítems para ser conservados en versiones futuras del instrumento.

Sin embargo, también se identificaron ítems con comportamientos anómalos o poco informativos, como **It_18**, con un valor atípicamente alto de discriminación, y **It_24** y **It_23**, con parámetros de dificultad extremadamente bajos o altos, respectivamente. Este tipo de resultados podría deberse a errores de codificación, baja frecuencia de respuesta o un desfase en la dificultad esperada del ítem en relación con la muestra analizada. En concordancia con Viladrich, Angulo-Brunet y Doval (2017), los parámetros extremos o negativos pueden reflejar inestabilidad en las estimaciones o problemas en el ajuste del modelo.

Desde un enfoque psicométrico, la presencia de ítems mal calibrados sugiere la necesidad de realizar ajustes o eliminar dichos reactivos, ya que su inclusión podría distorsionar la medición del rasgo latente y reducir la precisión diagnóstica del instrumento. Como señalan Oviedo y Campo-Arias (2005), un instrumento confiable y válido requiere ítems que no solo

discriminen adecuadamente, sino que también estén distribuidos en un rango de dificultad que refleje fielmente el continuo del constructo.

En términos generales, los hallazgos respaldan parcialmente la calidad del instrumento, al evidenciar la existencia de un conjunto de ítems funcionales. No obstante, se hace imprescindible un proceso de revisión y depuración adicional, centrado en los ítems con parámetros extremos, con el fin de asegurar una medición precisa, justa y útil en contextos educativos reales.

6.2 Contraste de los principales hallazgos con la teoría

Los hallazgos del presente estudio revelan discrepancias entre el modelo teórico propuesto para el instrumento y su comportamiento empírico. Aunque la validez de contenido fue confirmada a través del juicio de expertos, los resultados del análisis factorial exploratorio (AFE) no respaldaron una estructura factorial clara. Este resultado contrasta con lo señalado por Lloret-Segura et al. (2014), quienes indican que una estructura válida debe reflejarse en la agrupación coherente de los ítems conforme a los factores esperados. Además, las cargas factoriales cruzadas, negativas o inferiores al umbral mínimo sugieren limitaciones en la dimensionalidad del instrumento.

Respecto a la confiabilidad, el coeficiente alfa obtenido fue considerablemente bajo, lo que indica escasa consistencia interna entre los ítems. Por su parte, el coeficiente omega de McDonald alcanzó un valor de .67, el cual, si bien es ligeramente superior, solo puede considerarse aceptable en contextos exploratorios (Viladrich, Angulo-Brunet & Doval, 2017; Ventura-León & Caycho-Rodríguez, 2017). Estos resultados refuerzan la necesidad de una revisión estructural del instrumento, tal como lo propone McDonald (1999), cuando se evidencian inconsistencias psicométricas que afectan la validez y confiabilidad de las mediciones.

En cuanto al análisis mediante la Teoría de Respuesta al Ítem (TRI), los resultados obtenidos a partir del modelo 2PL evidenciaron una alta variabilidad en los parámetros de discriminación (α) y dificultad (β), incluyendo valores extremos o poco interpretables en varios ítems. Esta situación coincide con lo señalado por Viladrich, Angulo-Brunet y Doval (2017), quienes destacan que la estabilidad de los parámetros es un indicador fundamental de la calidad de los reactivos. Asimismo, la presencia de ítems con bajos niveles de

discriminación o dificultades atípicas refleja inconsistencias psicométricas que, como advierten Oviedo y Campo-Arias (2005), pueden comprometer la validez del instrumento al dificultar una medición precisa del constructo. Por ello, desde un enfoque psicométrico moderno, se recomienda revisar o eliminar aquellos ítems que no aportan información útil, a fin de optimizar la estructura del instrumento.

6.3 Implicaciones teóricas y prácticas de los resultados

Los hallazgos obtenidos en este estudio tienen implicaciones tanto teóricas como prácticas. Desde el plano teórico, la ausencia de una estructura factorial clara y la baja consistencia interna del instrumento sugieren la necesidad de revisar el modelo conceptual que lo sustenta, así como los componentes que definen la competencia lectora en inglés como lengua extranjera en el nivel educativo evaluado. Este resultado evidencia que la operacionalización del constructo podría requerir una mayor especificidad o una redefinición de sus dimensiones, lo cual aporta a la discusión académica sobre cómo abordar su medición en contextos educativos. En el ámbito práctico, se desprende que el instrumento, en su versión actual, no es apto para ser utilizado con fines evaluativos o diagnósticos, ya que podría inducir a interpretaciones erróneas sobre el nivel de desempeño de los estudiantes. Sin embargo, el proceso metodológico desarrollado sienta las bases para el rediseño del instrumento y enfatiza la necesidad de contar con herramientas válidas y confiables que respondan a las características del sistema educativo hondureño y a las demandas propias de la enseñanza del inglés como lengua extranjera.

Desde una perspectiva psicométrica, los resultados obtenidos mediante el modelo 2PL de la Teoría de Respuesta al Ítem (TRI) aportan evidencia valiosa sobre el comportamiento del instrumento diseñado para evaluar la competencia lectora en inglés como lengua extranjera (EFL). La estimación de los parámetros de discriminación (α) y dificultad (β) permitió comprender con mayor profundidad el funcionamiento de cada ítem dentro del continuo de habilidad lectora, superando los límites de la teoría clásica de los test. Estos hallazgos respaldan el uso de modelos de medición modernos en contextos educativos latinoamericanos, donde históricamente se ha recurrido a enfoques más tradicionales, y subrayan la necesidad de considerar la teoría del rasgo latente como base para el desarrollo de instrumentos más precisos y culturalmente adecuados. En el ámbito práctico, este análisis

permitió identificar ítems con buena capacidad discriminativa, así como reactivos que requieren ajustes por presentar parámetros inestables o extremos. Esta información resulta clave para la toma de decisiones pedagógicas, ya que facilita la depuración del instrumento y su eventual uso como herramienta diagnóstica en contextos escolares. En particular, una versión refinada del test podría implementarse para evaluar niveles de competencia lectora, apoyar programas de reforzamiento o monitorear el progreso de los estudiantes en el área de inglés como lengua extranjera, contribuyendo así a una evaluación más válida, confiable y contextualizada.

Conclusiones

La presente tesis tuvo como objetivo desarrollar un instrumento para evaluar la competencia lectora en inglés como lengua extranjera (EFL) en estudiantes del primer ciclo de educación básica en Honduras. Para ello, se llevó a cabo un proceso riguroso de construcción que incluyó la elaboración de ítems basados en criterios curriculares y lingüísticos, seguido de una serie de análisis psicométricos que abarcaron desde la validez de contenido mediante juicio de expertos hasta la aplicación de la Teoría de Respuesta al Ítem. Los resultados obtenidos proporcionan información valiosa para orientar futuras acciones por parte de la Secretaría de Educación en relación con la evaluación de la competencia lectora en inglés. Asimismo, este trabajo puede servir como referente metodológico para el diseño y construcción de instrumentos similares en contextos educativos equivalentes.

Una de las principales evidencias de validez obtenidas en este estudio fue la validez de contenido, evaluada a través del juicio de expertos. Los ítems contruidos fueron sometidos a revisión por parte de especialistas en el área de enseñanza del inglés y evaluación educativa, quienes valoraron aspectos como la claridad, pertinencia, redacción y congruencia con los objetivos del constructo. Los resultados mostraron niveles adecuados de acuerdo entre jueces, lo que indica que los ítems fueron considerados apropiados y representativos del dominio de la competencia lectora en inglés. Esta evidencia respalda la adecuación conceptual y lingüística del instrumento en la etapa inicial de construcción.

El análisis factorial exploratorio (AFE) realizado sobre el instrumento no confirmó una estructura factorial clara. Los modelos de uno, dos y tres factores no lograron evidenciar una agrupación consistente de los ítems, observándose cargas factoriales negativas y cruzadas en varias soluciones. Estos resultados sugieren que la organización interna del instrumento, en su forma actual, no refleja con precisión una dimensión subyacente única o claramente definida para evaluar la competencia lectora en inglés.

El análisis de confiabilidad evidenció niveles bajos de consistencia interna en el instrumento construido. El coeficiente alfa de Cronbach obtenido fue de .30, mientras que el coeficiente omega de McDonald alcanzó un valor de .67. Estos resultados indican que los ítems, en conjunto, no muestran una alta homogeneidad en sus respuestas, lo cual limita la

estabilidad interna de la escala para evaluar de manera consistente la competencia lectora en inglés en el grupo de estudiantes analizado.

El análisis efectuado mediante la Teoría de Respuesta al Ítem (TRI), específicamente a través del modelo logístico de dos parámetros (2PL), permitió cumplir con el objetivo de examinar el comportamiento psicométrico de los ítems del instrumento construido para evaluar la competencia lectora en inglés como lengua extranjera (EFL). La estimación de los parámetros de discriminación (α) y dificultad (δ) reveló la existencia de reactivos con un adecuado funcionamiento, capaces de diferenciar eficazmente entre distintos niveles de habilidad, así como ítems con parámetros atípicos que comprometen la calidad técnica del instrumento.

Estos hallazgos confirmaron que la TRI constituye una herramienta de análisis avanzada y pertinente, que aporta criterios objetivos para la toma de decisiones sobre la conservación, revisión o eliminación de ítems. En consecuencia, el uso del modelo 2PL permitió no solo identificar debilidades en la prueba, sino también ofrecer una base sólida para su futura optimización. En síntesis, los resultados obtenidos mediante la TRI fortalecen la evidencia psicométrica del instrumento en construcción y aportan lineamientos concretos para su mejora. La aplicación de este modelo en el contexto educativo hondureño representa un avance metodológico relevante, al introducir prácticas de evaluación más precisas, rigurosas y contextualizadas.

Recomendaciones

A partir de los resultados obtenidos en el análisis psicométrico del instrumento, se proponen las siguientes recomendaciones para su mejora y futura aplicación:

1. Se recomienda analizar si la prueba de inglés como lengua extranjera (EFL) puede ser utilizada como una herramienta diagnóstica útil en el entorno escolar. Esta revisión debe tomar en cuenta su capacidad para detectar con precisión el nivel de cada estudiante, así como su facilidad de aplicación dentro de las condiciones reales del sistema educativo.
2. Ante la falta de una estructura factorial clara, se aconseja revisar el modelo conceptual que sustenta el instrumento, de modo que se explore si la naturaleza del constructo requiere una estructura diferente (por ejemplo, más dimensiones, dimensiones jerárquicas, etc.).
3. Se recomienda aplicar el instrumento revisado a una muestra más grande y diversa, lo que permitiría mejorar la estabilidad de los análisis factoriales y aumentar la generalización de los resultados.
4. Se recomienda continuar con el proceso de validación del instrumento, reconociendo que la obtención de evidencias de validez es un procedimiento continuo y multidimensional (Elosua Oliden, 2003). En particular, se sugiere fortalecer las evidencias de validez relacionadas con el desempeño académico de los estudiantes en el contexto escolar. Este tipo de validación permitirá generar perfiles más precisos y confiables de los estudiantes que se encuentran en el segundo o tercer ciclo de educación primaria, facilitando así la planificación de estrategias pedagógicas que sienten bases sólidas para el aprendizaje del inglés como lengua extranjera. A largo plazo, esto podría traducirse en mayores oportunidades educativas y profesionales tanto a nivel nacional como internacional.
5. Además se recomienda realizar una revisión técnica exhaustiva de los ítems que presentaron parámetros atípicos de discriminación y dificultad en el modelo 2PL, con el fin de ajustar, reformular o eliminar aquellos reactivos que no contribuyen significativamente a la medición del constructo. Este proceso debe ir acompañado de una nueva fase de aplicación

piloto, que permita verificar la mejora en la calidad psicométrica del instrumento. Asimismo, se sugiere continuar empleando enfoques basados en la Teoría de Respuesta al Ítem en futuras investigaciones psicométricas, dada su capacidad para proporcionar información detallada sobre el comportamiento individual de cada ítem y su contribución al rasgo latente evaluado.

DEGT-UNAH

Referencias Bibliográficas

- Aiken, L. R. (2003). *Tests Psicologicos y Evaluacion undecima edicion*. Mexico: PEARSONEDUCACION.
- Alderson, J. C. (2000). *Assessing Reading*. Cambridge University Press.
- Alderson, J. C. (2018). *The politics of language education: Individuals and institutions*. Multilingual Matters.
- Aliaga Tovar, J. (2007). Psicometria: Tests Psicométricos, Confiabilidad y Validez. *Psicología: Tópicos de actualidad*, 85-108. doi:doi:https://d1wqtxts1xzle7.cloudfront.net/33465691
- Alonso, J. C. (2021). Nivel de inglés de los graduados de programas de Contaduría en Colombia: muy lejos de la meta. *Estudios Gerenciales*, 37(160), 335-348. doi:https://doi.org/10.18046/j.estger.2021.160.3878
- Amaya L. J. L. and Pacheco D., P. N. (2002). ANÁLISIS FACTORIAL DINÁMICO MEDIANTE EL MÉTODO TUCKER3. *Revista Colombiana de Estadística*, 25(1), 43-57. Obtenido de https://revistas.unal.edu.co/index.php/estad/article/view/28544
- American Educational Research Association, A. P. (2018). *Estándares para pruebas educativas y psicológicas (M. Lieve, Trans.)*. American Educational Research Association. (Trabajo original publicado en 2014). Obtenido de https://www.aera.net/Publications/Books/Standards-for-Educational-and-Psychological-Testing-2014-Edition
- Anastasi, A. &. (1998). *Psychological testing (7th ed.)*. Prentice Hall.
- APA, AERA, NCME. (2014). *Estandares para pruebas educativas y psicologicas (Spanish Edition)*. Washington, DC: American Educational Research Association. Edicin de Kindle. Original Work published 2014.
- Arthur, W. D. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, 56(1), 125-153.
- Ashton, K. J. (2012). *Technical report for the first European Survey on Language Competences*. Europe: European Comission.
- Assesment, E. L. (2025). *Cambridge University Press & Assessment 2025*. Obtenido de Cambridge University Press & Assessment 2025.

- Assessment, C. (19 de mayo de 2021). *Cambridge*. Obtenido de Cambridge: <https://www.cambridgeenglish.org/>
- Assessment, C. U. (2024). *English Language Assessment*. Obtenido de English Language Assessment: <https://www.cambridge.org>
- Attorresi, H. F. (2009). Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica* , 179-188.
- Aune, S. E. (2019). Análisis psicométrico mediante la Teoría de la Respuesta al Ítem: Su utilización en America Latina. Supuestos de unidimensionalidad e independencia local. *International Journal of Psychological Research*, 12(1), 49-56. Obtenido de <http://doi.org/10.21500/20112084.3753>
- Baker, F. B. (2001). *The basics of item response theory (2nd ed.)*. ERIC Clearinghouse on Assessment and Evaluation.
- Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer.
- Bartlett. (1954). A Note on the Multiplying Factors for Various Chi Square Approximations. *Journal of the Royal Statistical Society*, 16(2), 296–298.
- Bazaga, F., & Marín, F. (2003). *AEL: Test for the Assessment of the English Language*. Madrid: TEA Ediciones.
- Beltran, M. (2017). El aprendizaje del idioma inglés como lengua extranjera. *Boletín Redipe*, 91-98.
- Beltrán-Santoyo, G. R.-H.-B. (2021). La importancia e influencia del idioma inglés dentro del campo científico. *Revista Lengua y Cultura*, 46-51. Obtenido de <https://repository.uaeh.edu.mx/revistas/index.php/lc/issue/archive>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (págs. (pp. 395-479).). Addison-Wesley.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74.
- British Council. (2023). *IELTS (International English Language Testing System)*. Obtenido de IELTS (International English Language Testing System): <https://www.britishcouncil.org>

- Brito-Chiquillo, J. G.-P.-P. (2023). Adopción del Inglés como Segunda Lengua, un Componente de Competitividad en la Educación. *Colombian Applied Linguistics Journal*, 25(1), 86-96. doi:<https://doi.org/10.14483/22487085.19003>
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
- Brookhart, S. M. (2019). *Educational assessment of students (8th ed.)*. Pearson.
- Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research (2nd ed.)*. The Guilford Press.
- Bruzual, R. (2007). Fundamentos teóricos y metodológicos para la enseñanza de la lengua materna (L1) y segundas lenguas (L2) en contextos bilingües. *Argos*, 24(46) 46-65. Obtenido de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0254-16372007000100006&lng=es&tlng=es.
- Burga Leon, A. (2006). La unidimensionalidad de un instrumento de medición: perspectiva factorial. *Revista de Psicología*, 53-80. Obtenido de <http://www.redalyc.org/articulo.oa?id=337829536003>
- Cambridge Assessment English. (2021). *Young Learners English (YLE) qualifications*. Obtenido de <https://www.cambridgeenglish.org>
- Cambridge Assessment. (2024). *B1 Preliminary (PET) Exam Overview*. Obtenido de <https://www.cambridgeenglish.org/exams-and-tests/preliminary/>
- Cangur, S. y. (2015). Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality. *Journal of Modern Applied Statistical Methods*, 14(1):152-167. doi::10.22237/jmasm/1430453580
- Caracas Sánchez, B. P. (2019). La evaluación de la comprensión lectora en México. El caso de las pruebas EXCALE, PLANEA y PISA. *Perfiles educativos*, 41(164), 8-27. Obtenido de <https://doi.org/10.22201/iisue.24486167e.2019.164.59087>
- Castillero Mimenza, O. (9 de Septiembre de 2024). *Psicología y Mente*. Obtenido de Psicología y Mente: <https://psicologiaymente.com/miscelanea/tipos-de-validez>
- Català i Agràs, G., Català Agràs, M., Molina i Hita, E., & Monclús i Bareche, R. (2001). Evaluación de la comprensión lectora : pruebas ACL (primero-sexto de primaria). *Redined: Red de Informacion Educativa*. Obtenido de <http://hdl.handle.net/11162/59795>

- Certified Test Center. (2023). *TOEFL, TOEIC y sus diferencias en la evaluación del inglés*. Obtenido de TOEFL, TOEIC y sus diferencias en la evaluación del inglés: <https://www.certifiedtestcenter.com>
- Cizek, G. J. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, 70(5), 732–743. doi:<https://doi.org/10.1177/0013164410379323>
- College Board. (2020). *College Board*. Obtenido de College Board: <https://collegereadiness.collegeboard.org/sat>
- Cortada de Kohan, N. (2002). Importancia de la investigación psicométrica. *Revista Latinoamericana de Psicología*, 229-240. Obtenido de <http://www.redalyc.org/articulo.oa?id=80534303>
- Cortes Velandia, G. F. (2018). *Creación de talleres en la plataforma moodle e implementación de tutorías virtuales en facebook para la preparación en la prueba saber 11 en el área de inglés*. Colombia: Licenciatura en Educación Básica con Énfasis en Inglés. Obtenido de <http://hdl.handle.net/11349/12883>
- Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3), 297–334.
- Chamorro-Ortega, C. P.-C. (2020). ¿Qué habilidad lingüística se hace más compleja enseñar en el aula de inglés. *Revista Científica Dominio de las Ciencias*, 302-318. Obtenido de <http://dominiodelasciencias.com/ojs/index.php/es/index>
- Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Jossey-Bass.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Press.
- Demarchi. (2020). La evaluación desde las pruebas estandarizadas en la educación en Latinoamérica. *Revista En-Contexto*, 107-133.
- Didier. (2019). Las pruebas estandarizadas: Cuando el neoliberalismo se sienta en las aulas. *Encuentros. Revista De Ciencias Humanas, Teoría*, Disponible en Redalyc.
- Digeduca. (2015). Análisis de Items con JMetrik : serie de cuadernillos técnicos. Guatemala. 1-31. Obtenido de https://issuu.com/digeduca/docs/analisis_items_jmetrik
- Dorans, C. &. (2016). *Fairness in educational assessment and measurement*. Routledge. Obtenido de

- https://www.researchgate.net/publication/364749881_Fairness_in_Educational_Assessment_and_Measurement
- Downing, S. M. (2006). *Handbook of test development*. Routledge.
- Duolingo. (10 de Marzo de 2023). *Duolingo English Test*. Obtenido de Duolingo English Test: <https://englishtest.duolingo.com/scores>
- Educación), I. (. (2018). *PLANEA: Evaluaciones del aprendizaje en México*. Mexico.
- Educación, S. d. (2020). *Informe estadísticas educativas de Honduras*. Honduras: Sistema de Administración de Centros Educativos (SACE,2020). Obtenido de <https://sace.se.gob.hn/>
- EF Education First. (2023). *EF English Proficiency Index 2023*. Recuperado de <https://www.ef.com/epi>.
- EF SET. (16 de February de 2025). *EF Standard English Test (EF SET)*. Obtenido de EF Standard English Test (EF SET): <https://www.efset.org/>
- Ehri, L. C. (2005). Development of Sight Word Reading: Phases and Findings. *The science of reading: A handbook*, 135-154. Obtenido de <https://doi.org/10.1002/9780470757642.ch8>
- Elosua Oliden, P. (2003). Sobre la validez de los test. *Psicothema*, 315-321. Obtenido de <http://www.redalyc.org/articulo.oa?id=72715225>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- English., Cambridge Assessment. (2021). *Young Learners English (YLE) qualifications*. Obtenido de <https://www.cambridgeenglish.org>
- Escobar Pérez, J., & Cuervo Martínez, A. (2008). Validez de contenido y juicio de expertos una aproximación a su utilización. *Avances en medicion*, 27-36.
- ETS. (18 de febrero de 2020). *ETS*. Obtenido de ETS: https://www.ets.org/toefl_primary
- ETS. (2021). *ETS*. Obtenido de ETS: <https://www.ets.org/toefl>
- ETS. (2021). *TOEFL® Research Insight Series, Volume 10: Using TOEFL iBT® Test Scores for Selecting International Teaching Assistants*. Educational Testing Service. Obtenido de <https://www.ets.org/toefl/research/insight-series>
- Faulkner-Bond, M. &. (2016). *Educational measurement: From foundations to future*. Guilford Press. Obtenido de <https://www.guilford.com/books/Educational-Measurement/Wells-Faulkner-Bond/9781462525621>

- FEREMA. (2017). *Informe del Progreso Educativo Honduras* . Obtenido de <https://issuu.com/ferema/docs/informe-de-progreso-educativo-web>
- Fernández Liporace, M., Ongarato, P., Saavedra, E., & Casullo, M. M. (2004). El Test de Matrices Progresivas, Escala General: un análisis psicométrico. *Revista Evaluar*, 50-69.
- Ferrando, P. J. (1966). Evaluación de la unidimensionalidad de los items mediante análisis factorial. *Psicothema*, 397-410. Obtenido de <http://www.redalyc.org/articulo.oa?id=72780215>
- Ferrando, P. J.-C. (2010). El análisis factorial como técnica de investigación en psicología [Factor analysis as a technique in psychological research]. *Papeles del Psicólogo*, 31(1), 18–33.
- Ficayuma, L. A. (2024). Genre analysis and textual boundaries to cultivate literacy and to formulate language English policy: A corpus study of PISA, TOEFL, English book and curriculum. *Proceedings of the ICONELT - International Conference on English Language Teaching.*, 1-15. Obtenido de e (<http://creativecommons.org/licenses/by-nc/4.0/>)
- Fraillon, J. A. (2019). *CILS 2018: Preparing for life in a digital world*. Springer.
- Freiberg Hoffmann, A. S. (2013). CORRELACIONES POLICÓRICAS Y TETRACÓRICAS EN ESTUDIOS FACTORIALES EXPLORATORIOS Y CONFIRMATORIOS. *Ciencias Psicológicas*, 7(2), 151-164. Obtenido de http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-42212013000200005&lng=es&tlng=es.
- Frias-Navarro, D. (2022). Apuntes de estimación de la fiabilidad de consistencia interna de los items de un instrumento de medida. 1-22. Obtenido de <https://www.uv.es/friasnav/AlfaCronbach.pdf>
- Galicia Alarcón, L. A. (2017). Validez de contenido por juicio de expertos: propuesta de una herramienta virtual. *Apertura (Guadalajara, Jal.)*, 9(2), 42-53. Obtenido de <https://doi.org/10.32870/ap.v9n2.993>
- Garcia, E. (1993). *Introduccion a la Psicometria*. España: Siglo XXI.
- Garduño Teliz, E. (2019). La Evaluacion Estandarizada del Desempeño Academico. *Research Gate*, 152.

- Gierl, M. J. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116. Obtenido de <https://doi.org/10.3102/0034654317726529>
- Gil-Gómez de Liaño, B. y.-E. (2019). La Metodología Delphi como tecnica de estudio de la validez de contenido. *Anales de Psicología*, 28(3). Obtenido de <http://dx.doi.org/10.6018/analesps.28.3.156211>
- Gómez, P., & Joan de, M.-A. J. (2009). Validación de la versión española del cuestionario sobre la práctica basada en la evidencia en enfermería. *Revista Española de Salud Pública*, 83(4), 577-586. Obtenido de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57272009000400009&lng=es&tlng=es
- Gonzales Acosta, M. (29 de Noviembre de 2019). *Miros*. Obtenido de Miros: <https://www.miros.ec/blog/2018/11/29/pruebas-estandarizadas-ventajas-y-desventajas/melvis/>
- Grabe, W., & Stoller, F. L. (2002). *"Teaching and Researching Reading"*. Longman.
- Guarín, M. C. (2017). Desarrollo de habilidades de comprensión lectora en inglés –como lengua extranjera– en estudiantes de quinto de primaria. *Revista Interamericana de Investigación, Educación y Pedagogía*, 10(2), 59-78. Obtenido de <https://www.redalyc.org/articulo.oa?id=561059354006>
- Güler, N. K. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *International Asociation of Social Science Research*, 2(1), 1-6.
- Gutiérrez Duque, D. M., & Mayora Pernía, C. A. (2021). Variables predictoras del desempeño escolar en exámenes estandarizados de inglés: Evidencias desde el examen de Estado en Colombia. *Revista Virtual Universidad Catolica del Norte*, 33-62. doi: <https://doi.org/10.35575/rvucn.n62a3>
- Haladyna, T. M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates.
- Hambleton, R. K. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing.
- Hambleton, R. K. (1991). *Fundamentals of item response theory*. SAGE.

- Hambleton, R. K. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Harden, R. M. (2016). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1(5955), 447-451.
- Harlen, W. (2007). *Assessment of learning*. Sage Publications.
- Hernandez-Nieto, R. (2002). *Contributions to Statistical Analysis: The Coefficients of Proportional Variance, Content Validity and Kappa*. CreateSpace Independent Publishing Platform. doi:1588987159, 9781588987150
- Hernandez-Sampieri, R. F.-C.-L. (2014). *Metodología de la Investigación 6ta edición*. Mexico: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.
- Hidalgo-Montesinos, M. D. (2016). Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas. *Revista de Psicología Clínica con Niños y Adolescentes [en línea]*, 3(2), 13-21[fecha de Consulta 3 de Octubre de 2024]. Obtenido de <https://www.redalyc.org/articulo.oa?id=477152554002>
- Höl, F. &. (2023). Standardized English language proficiency tests for young learners: An overview. In Current Studies in Foreign Language Education. *International Society for Research in Education and Science (ISRES)*, 346-359. Obtenido de <https://www.isres.org/>
- ICFES. (2020). *Pruebas SABER: Evaluaciones del sistema educativo colombiano*. Colombia.
- IEA. (16 de febrero de 2025). *IEA: Progress in International Reading Literacy Study (PIRLS): Evaluación internacional de comprensión lectora en inglés*. Obtenido de IEA: Progress in International Reading Literacy Study (PIRLS): Evaluación internacional de comprensión lectora en inglés: <https://www.iea.nl/studies/iea/pirls>
- (INEE), I. N. (2019). *Guía para la elaboración de instrumentos de evaluación*. Mexico: INEE. Obtenido de <https://www.inee.edu.mx/wp-content/uploads/2019/04/P1E213.pdf>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. Obtenido de <https://doi.org/10.1098/rsta.2015.0202>
- Kaiser. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. Obtenido de <https://doi.org/10.1007/BF02291575>

- Klein, S. B. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, , 31(5), 415-439.
- Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.
- Ledesma, R. D. (2019). Uso del Análisis Factorial Exploratorio en RIDEP. Recomendaciones para Autores y Revisores [The use of exploratory factor analysis in RIDEP. Guidelines for authors and reviewers]. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica*, 52(3), 173–180. Obtenido de <https://doi.org/10.21865/RIDEP52.3.13>
- Lira, M. M. (2024). Procesos Evaluativos y Pruebas Estandarizadas: ¿Son compatibles si buscamos la calidad en la educación? *Region Científica*, 1-11. Obtenido de <https://doi.org/10.58763/rc2024204>
- Llorens, A. (2015). Estrategias para la enseñanza de la comprensión lectora. *Revista de Educación y Psicopedagogía*, 12(1), 45-60.
- Lloret-Segura, S. F.-T.-B.-M. (2014). El Análisis Factorial Exploratorio de los Ítems: una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), 1151-1169. Obtenido de <https://dx.doi.org/10.6018/analesps.30.3.199361>
- Lopez, J. A. (1995). Estimacion de parametros en la TRI: Una evaluacion de BILOG en muestras pequeñas. . *Psicothema*, 7(1), 173-185. Obtenido de <http://www.redalyc.org/articulo.oa?id=72770115>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- LPI-SEDUC-001-2015. (2015). Estándares Educativos Nacionales-Ingles. Comayagua: Impresión y distribución de Textos y Adquisición de Materiales Escolares en el Marco de la Propuesta Curricular del Plan EFA-2015.
- Marco, I. T., González Roma, V., & Gómez Benito, J. (2000). Teoría de respuesta al ítem y análisis factorial confirmatorio: dos métodos para analizar la equivalencia psicométrica en la traducción de cuestionarios. *Psicothema*, 540-544. Obtenido de <http://www.redalyc.org/articulo.oa?id=72797122>
- Martinez Arias, M. R., Hernández Lloreda, M. V., & Hernández Lloreda, M. J. (2006). *Psicometria*. Madrid: Alianza Editorial, S. A. Obtenido de www.alianzaeditorial.es
- Matas Terron, A. (2010). *Introducción al análisis de la Teoría de Respuesta al Ítem*.

- Mavrou, I. (2015). Análisis factorial exploratorio Cuestiones conceptuales y metodológicas. *Revista Nebrija De Lingüística Aplicada a La Enseñanza De Lenguas*, (19), 71–80. Obtenido de <https://doi.org/10.26378/rnlael019283>
- McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates.
- McMillan, J. H. (2015). *Classroom assessment: Principles and practice for effective standards-based instruction (7th ed.)*. Pearson.
- McNamara, D. S. (2004). *Reading comprehension strategies: Theories, interventions, and technologies*. Lawrence Erlbaum Associates.
- MECD. (2012). *Estudio europeo de Competencia Linguistica EECL*.
- Medrano, L., & Pérez, E. (2020). *Manual de psicometría y evaluación psicologica*. Cordoba: Facultad de Psicología. Obtenido de <http://hdl.handle.net/11086/21769>
- Mejia Obando, K. A. (2016). Uso de canciones en educación media para la enseñanza del idioma inglés en Tegucigalpa, Honduras. 90p. Obtenido de <https://repositorio.ufscar.br/handle/ufscar/8677>
- Melia, J.L. (1990). *La construccion de la psicometria como ciencia teorica y aplicada*. Valencia: Cristobal Serrano. Obtenido de <http://www.uv.es/psicometria>
- Mendoza Navas, B. &. (2024). Validez, confiabilidad y amenazas a la validez en las evaluaciones de la clase de Inglés I virtual del Departamento de Lenguas Extranjeras de la Universidad Nacional Autónoma de Honduras. *MLS Educational Research (MLSER)*, 9(1). Obtenido de <https://doi.org/10.29314/mlser.v9i1.2523>
- MEP. (3 de junio de 2024). *Ministerio de educacion publica*. Obtenido de Ministerio de educacion publica: <https://www.mep.go.cr/>
- MINEDU. (29 de noviembre de 2024). *El Pais*. Obtenido de El Pais: <https://elpais.com/>
- Ministerio de educacion, formacion profesional y deportes. (2023). *Programa para la Evaluación Internacional de los Estudiantes: PISA 2022, Informe español*. Madrid: Instituto Nacional de Evaluación Educativa. Obtenido de <https://cpage.mpr.gob.es/>
- Montero Rojas, E. (2013). Referentes Conceptuales y metodologicos sobre la nocion moderna de validez de instrumentos de medicion: implicaciones para el caso de personas con necesidades educativas especiales. . *Actualidades En Psicologia*, 27(114), 113-128.
- Moore, D. P. (2025). Try, translate, relate: A student tool for remixing English language arts assessment prompts. *International Journal of Student Voice*, 11(2). Obtenido de <https://sites.psu.edu/ijsv/volume-11/>

- Morales Vallejo, P. (2007). La fiabilidad de los tests y escalas. 36p. Obtenido de <https://studylib.es/doc/5105328/la-fiabilidad-de-los-tests-y-escalas>
- Morales-Ocaña, A., & Higuera-Rodríguez, M. L. (2017). PROCESOS DE ENSEÑANZA-APRENDIZAJE. ESTUDIOS, AVANCES Y EXPERIENCIAS. *Revista de Currículo y Formación de Profesorado*, 1-6. Obtenido de <http://www.redalyc.org/articulo.oa?id=56752038001>
- Mullis, I. V. (2017). *PIRLS 2016 international results in reading*. IEA.
- Mullis, I. V. (2019). *TIMSS 2019 international results in mathematics and science*. IEA.
- Mullis, I. V. (2023). *PIRLS 2016 International Results in Reading*. Boston: TIMSS & PIRLS International Study Center. Obtenido de <https://doi.org/10.6017/lse.tpisc.tr2103.kb5342>
- Muñiz. (2010). Las teorías de los test: Teoría clásica y Teoría de respuesta a los ítems test theories: classical theory and item response theory. *Papeles del Psicólogo*, 57-66. Obtenido de <http://www.cop.es/papeles>
- Muñiz. (2018). *Introducción a la Psicometría Teoría Clásica y TRI*. Madrid: Ediciones Pirámide (Grupo Anaya, S. A.).
- Muñiz, & Fonseca - Pedrero. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema*, 31(1), 7–16. Obtenido de <https://doi.org/10.7334/psicothema2018.291>
- Navas Ara, M. J. (2001). *Métodos, diseños y técnicas de investigación psicológica*. España. Obtenido de <https://dialnet.unirioja.es/servlet/libro?codigo=325189>
- Navas Ara, M. J. (2002). *Psicothema*. Madrid: UNED.
- Navas, A. M. (2001). Métodos, Diseños y Técnicas de Investigación Psicológica. *Psicothema*, 14(4), 871.
- Nitko, A. J. (2011). *Educational assessment of students (6th ed.)*. Pearson.
- Núñez, K. L. (29 de julio de 2024). *Bridge Universe*. Obtenido de Bridge Universe: <https://bridge.edu/tefl/blog/que-es-el-mcer-el-gse-niveles-de-dominio-del-ingles/>
- OECD. (2018). *TALIS 2018 results: Teachers and school leaders as lifelong learners*. OECD Publishing.
- OECD. (2019). *Skills Strategy 2019: Skills to shape a better future*. OECD Publishing. Obtenido de https://www.oecd.org/content/dam/oecd/es/publications/reports/2019/05/oecd-skills-strategy-2019_g1g9ff20/e3527cfb-es.pdf

- Ortiz, M. (2014). Introducción a la historia y conceptos básicos de la psicometría. 1-13. Obtenido de <https://docplayer.es/20398067-Introduccion-a-la-historia-y-conceptosbasicos-de-la-psicometria-melissa-judith-ortiz-barrero-mg-psicologia.htm>
- Oviedo, H. C., & Campo-Arias, A. . (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista Colombiana de Psiquiatría*, 34(4), 572–580.
- Pais, E. (29 de noviembre de 2024). *El Pais*. Obtenido de El Pais: <https://elpais.com/>
- Pedrosa, I. S.-Á.-C. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica*, 10(2), 3-18. doi:<https://dx.doi.org/10.5944/ap.10.2.11820>
- Pedrosa, I. S.-Á.-C. (2014). Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación. *Acción Psicológica*, 10(2), 3–18. doi:<https://doi.org/10.5944/ap.10.2.11820>
- Pellegrino, J. W. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
- Perassi, Z. (2008). *La Evaluacion en Educacion: Un campo de controversias*. Argentina: Ediciones LAE Laboratorio de Alternativas Educativas.
- Pérez de Pérez, A. &. (2006). La prueba de aptitud académica: una visión de la subprueba de comprensión de lectura. *Investigación y Postgrado*, 21(2), 143-176. Obtenido de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-00872006000200006&lng=es&tlng=es
- Pérez, E. &. (2010). Análisis factorial exploratorio Bases conceptuales y metodológicas. *Revista Argentina de Ciencias del Comportamiento (RACC)*, 58-66.
- Popham, W. J. (2017). *Classroom assessment: What teachers need to know (8th ed.)*. Pearson.
- Prieto, G. &. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, 31(1), 67-74.
- Prisbey, C. A. (2013). Bilinguismo. *Revista digital de FEAE-Aragón sobre organización y gestión educativa*, 19-20. Obtenido de <https://dialnet.unirioja.es/servlet/revista?codigo=16455>
- Ramírez, L. (2017). Niveles de comprensión lectora en inglés como lengua extranjera. *Revista de Lingüística Aplicada*, 15(2), 120-138.
- Ramos, L., & Casas, L. M. (2018). Enseñanza y evaluación del Álgebra en Honduras: concepciones y prácticas docentes. *Paradigma. Revista de investigación educativa*, 24(38):34. doi:[10.5377/paradigma.v24i38.6769](https://doi.org/10.5377/paradigma.v24i38.6769)

- Ravela, P. A. (2008). Las Evaluaciones Educativas que America Latina necesita. *Revista Iberoamericana de Evaluacion Educativa*, 46-63. Obtenido de <http://www.rinace.net/riee/numeros/vol1-num1/art4.pdf>
- Reise, S. P., & Revicki, D. A. . (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.
- Reyes Vásquez, A. (2023). Evaluación de la comprensión lectora en inglés en estudiantes del tercer ciclo de educación básica en Honduras. *Revista Paradigma.*, Recuperado de <https://camjol.info>.
- Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 108-116. Obtenido de <http://www.redalyc.org/articulo.oa?id=72729538017>
- Robitzsch, A. K. (2020). *TAM: Test analysis modules (Versión 3.5-19) PaquetedeR*. CRAN. Obtenido de <https://CRAN.R-project.org/package=TAM>
- Robles Garrote, P. y. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. *Revista Nebrija de Lingüística Aplicada en la enseñanza de las lenguas*, 18.
- Rojas, L. (2014). Evidencias de validez de la Prueba de Aptitud Académica de la Universidad de Costa Rica basadas en su estructura interna. *Actualidades en Psicología*, 15-26.
- Rueda Cataño, M. C., & Wilburn Dieste, M. (2014). Enfoques teóricos para la adquisición de una segunda lengua desde el horizonte de la practica educativa. *Perfiles Educativos*, 36(143), 21-28. Obtenido de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982014000100018&lng=es&tlng=es.
- Sanchez- Jabba, A. (2013). Bilinguismo en Colombia. *Centro de Estudios Economicos Regionales*.
- Sánchez, D. G. (2020). La evaluación desde las pruebas estandarizadas en la educacion en Latinoamerica. *En-Contexto Revista de Investigación en Administración, Contabilidad, Economía y Sociedad*, 107-121. Obtenido de Redalyc: <https://www.redalyc.org/articulo.oa?>
- Sanchez, R. (2021). El tema de validez de contenido en la educación y la propuesta de Hernandez-Nieto. *Lat. Am. J. Phys. Educ.*, 1-5. Obtenido de <http://www.lajpe.org>

- Scholastica Prep. (10 de Marzo de 2023). *Duolingo English Test: La alternativa al TOEFL y IELTS*. Obtenido de Duolingo English Test: La alternativa al TOEFL y IELTS: <https://www.scholasticaprep.com>
- Secretaría de Educación de Honduras. (2021). *Informe sobre el estado de la educación en Honduras*. Tegucigalpa, Honduras.
- SET, E. (2016). *EF SET Certificate*. Obtenido de EF SET Certificate: <https://www.efset.org>
- Shuttleworth, M. (23 de Octubre de 2020). *Explorable, Think outside the box*. Obtenido de Explorable, Think outside the box: <https://explorable.com/es/tipos-de-validez>
- Smith, G. M. (2003). Theoretical and empirical advances in construct validation. *Psychological Assessment*, 15(4), 455–467. Obtenido de <https://doi.org/10.1037/1040-3590.15.4.455>
- Soriano Rodriguez, A. M. (2014). Diseño y validación de instrumentos de medición. 19-40.
- Stiggins, R. J. (2008). *Assessment for learning: Doing it right, using it well*. Pearson.
- Susan E. Embretson, S. P. (2000). *Item Response Theory for Psychologists*. New York: Psychology Press. Obtenido de <https://doi.org/10.4324/9781410605269>
- Tajik, A. (2025). Integrating AI-Driven Emotional Intelligence in Language Learning Platforms to Improve English Speaking Skills through Real-Time Adaptive Feedback. *Research Square*, 40. Obtenido de <https://orcid.org/0000-0002-4352-5523>
- Tarrant County College. (2023). *ESL Placement Test: English as a Second Language Exam*. Obtenido de ESL Placement Test: English as a Second Language Exam.: <https://www.tcc.edu>
- Taylor, L. (2013). *Validity and validation in language testing*. . Cambridge University Press.
- Thorne, C. R. (1988). La Escala de Inteligencia para Niños de Wechsler y las variantes socio-culturales. *Revista de Psicología*, 51-62. doi: <https://doi.org/10.18800/psico.198801-02.004>
- Toland, M. D. (2013). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of early Adolescence*, 34(1):120-151. doi:10.1177/0272431613511332
- <https://www.cambridgeenglish.org/exams-and-tests/preliminary/>. (2024). *1 Preliminary (PET) Exam Overview*. Obtenido de <https://www.cambridgeenglish.org/exams-and-tests/preliminary/>
- UNESCO. (2019). *Estudio Regional Comparativo y Explicativo (ERCE) 2019: Resultados de aprendizaje en América Latina y el Caribe*. United Nations Educational, Scientific and Cultural Organization. Obtenido de <https://www.unesco.org>

- UNESCO. (2021). Los aprendizajes fundamentales en America Latina y el Caribe; evaluacion de logro de los estudiantes:estudioregional comparativo y explicativo(ERCE 2019)Resumen Ejecutivo. *UNESDOC, Biblioteca Digital*. Obtenido de <https://unesdoc.unesco.org/>
- Unesco Publishing. (2010). *Atlas of the World's Languages in Danger*. France: United Nations Educational,Scientific and Cultural Organization. Obtenido de atlas@unesco.org
- UPAEP. (16 de Febrero de 2025). *UPAEP IDIOMAS*. Obtenido de UPAEP IDIOMAS: <https://upaep.mx/>
- Urrutia Egaña, M. B. (2014). Métodos óptimos para determinar validez de contenido. *Educación Médica Superior*, 28(3), 547-558. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412014000300014&lng=es&tlng=es
- Van der Linden, W. J. (2010). *Elements of adaptive testing*. Springer.
- Van der Linden, W. J., & Hambleton, R. K. . (2013). *Handbook of modern item response theory*. Springer.
- Van Teun Adrianus, D. I. (s.f.). *Strategies of Discourse Comprehension*. Nueva York: Academic Press.
- Ventura-Leon, J. L. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, 625-627.
- Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología*, 33(3), 755–782. Obtenido de <https://doi.org/10.6018/analesps.33.3.268401>
- Villardón, J. L. (2007). Introducción al análisis de clúster. 22p.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.
- Wolf, M. K. (2024). Design Framework for the TOEFL Primary® Writing Test. *Educational Testing Service (ETS)*, Research Memorandum No. RM-24-02. Obtenido de ETS International English Language Testing System: <https://www.ets.org/contact/additional/research.html>

Yusof, M. (2023). The six stages of test construction. *ResearchGate*. Obtenido de https://www.researchgate.net/publication/374782421_The_Six_Stages_of_Test_Construction

DEGT-UNAH

Declaración de Ética de Investigación

Esta investigación ha sido realizada tomando en consideración el código de ética de investigación. No se reportan conflictos de interés.

DEGT-UNAH

Anexos

Anexo A

Tabla de Especificaciones de la prueba de inglés como lengua extranjera (EFL) que evalúa la competencia lectora

Block	Component	Standard	# items	Score (%)
Communication	Communicate in another language	1. Relate the sequence of a story by following the illustrations	17	20%
		2. Read and respond to a variety of texts (designing a poster, making a graph, etc).		
Culture	Gain knowledge and understanding of other cultures.	3. Read and identify the main idea in authentic texts.	11	20%
		4. Write a personal letter with key words (e.g., Dear..., Sincerely, Yours Truly, Love, etc).		
Culture	Gain knowledge and understanding of other cultures.	5. Engage in the writing and discussion of stories.	11	20%
		1. Read and respond to simple comprehension questions information found in cookbooks, folktales, and short stories on a variety of cultural holidays and festivals.		
Culture	Gain knowledge and understanding of other cultures.	2. Describe natural-ecological, cultural, archeological and engineering achievements, and/or symbols of the Honduran culture and the target culture.	11	20%

Connections	Connect with other disciplines and acquire information.	1. Classify foods according to colors, shapes and food groups. 2. Make connections between months, seasons and national holidays. 3. Use the internet, newspaper, or other sources, to read, take notes or record and graph Information (e.g., the weather in capital city or another town of target country for 1 week). 4. Match story sequence read aloud to a series of pictures (e.g., "Once upon a time...and they lived happily ever after."). 5. Recall persons (characters) or settings of stories from picture books.	10	30%
Comparisons	Develop insight into the nature of language and culture.	1. Compare weather and the implications (e.g., dress, sports, foods, houses) in the Honduran culture with the target culture.	4	20%
Communities	Participate in multilingual communities at home and around the world.	1. Research and list holidays and special celebrations from the target culture.	8	10%

Nota: Elaboracion propia

Anexo B

Carta de autorización de la institución educativa para la aplicación del instrumento



UNAH
UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS



**MAESTRÍA EN
PSICOMETRÍA Y
EVALUACIÓN EDUCATIVA**

cienciasociales.unah.edu.hn
mpse@unah.edu.hn
Tel: (504) 2216-6100 Ext. 100470

Tegucigalpa, M. D.C. 10 de Mayo de 2024

Señores Coordinación Académica del C. E. B. Juan Ramón Molina

Por medio del presente me dirijo a usted para solicitar su autorización para aplicar un instrumento de investigación en los estudiantes de cuarto grado de primaria de la institución que usted dignamente dirige. Dicha actividad forma parte fundamental del desarrollo de mi proyecto de tesis titulado **“Validación de una prueba de inglés EFL) que permita evaluar la competencia lectora en estudiantes de primer ciclo de educación primaria”**, requisito indispensable para obtener el grado en **Maestría en Psicometría y Evaluación Educativa** que pertenece a la Universidad Nacional Autónoma de Honduras.

El instrumento mencionado tiene como objetivo evaluar la competencia lectora en inglés como lengua extranjera (EFL) en los estudiantes de cuarto grado de primaria. Es importante mencionar que la información recopilada será utilizada estrictamente con fines académicos, manteniendo la confidencialidad y anonimato de todos los participantes involucrados.

Quedo a su disposición para cualquier aclaración o información adicional que considere pertinente y agradezco de antemano su valiosa colaboración y apoyo en este proceso académico.

Atentamente,

Suany Lisbeth Calderon Ruiz Cuenta No. PSE100208

slisbeth1984@gmail.com celular: 32072922

Por medio de la presente, la coordinación académica, autoriza formalmente a la estudiante Suany Calderon a aplicar el instrumento de evaluación dentro de nuestra institución educativa. Esta autorización se concede exclusivamente para fines académicos, específicamente como parte del desarrollo de su formación profesional en la **Maestría en Psicometría y Evaluación Educativa**. Se establece claramente que la información obtenida será manejada de manera estrictamente confidencial y utilizada únicamente con propósitos académicos y científicos.


Firma y sello Coordinación Académica

Anexo C

Carta de autorización de la institución educativa para la aplicación del instrumento







**MAESTRÍA EN
PSICOMETRÍA Y
EVALUACIÓN EDUCATIVA**

cienciasociales.unah.edu.hn
 mpse@unah.edu.hn
 Tel: (504) 2216-6100 Ext. 100470

Tegucigalpa, M. D.C. 10 de Mayo de 2024

Señores Directores del
instituto CEB 4 de Junio

Estimado(a) director(a) Eduifo Moradel Arcona

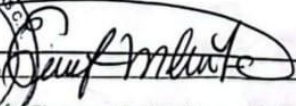
Por medio del presente me dirijo a usted para solicitar su autorización para aplicar un instrumento de investigación en los estudiantes de cuarto grado de primaria de la institución que usted dignamente dirige. Dicha actividad forma parte fundamental del desarrollo de mi proyecto de tesis titulado **“Validación de una prueba de Inglés EFL) que permita evaluar la competencia lectora en estudiantes de primer ciclo de educación primaria”**, requisito indispensable para obtener el grado en Maestría en Psicometría y Evaluación Educativa que pertenece a la Universidad Nacional Autónoma de Honduras.

El instrumento mencionado tiene como objetivo evaluar la competencia lectora en inglés como lengua extranjera (EFL) en los estudiantes de cuarto grado de primaria. Es importante mencionar que la información recopilada será utilizada estrictamente con fines académicos, manteniendo la confidencialidad y anonimato de todos los participantes involucrados.

Quedo a su disposición para cualquier aclaración o información adicional que considere pertinente y agradezco de antemano su valiosa colaboración y apoyo en este proceso académico.

Aientamente,
 Suany Lisbeth Calderon Ruiz Cuenta No. PSE100208
slisbeth1984@gmail.com celular: 32072922

Por medio de la presente, el suscrito Director del Centro Educativo, autoriza formalmente a la estudiante Suany Calderon a aplicar el instrumento de evaluación dentro de nuestra institución educativa. Esta autorización se concede exclusivamente para fines académicos, específicamente como parte del desarrollo de su formación profesional en la **Maestría en Psicometría y Evaluación Educativa**. Se establece claramente que la información obtenida será manejada de manera estrictamente confidencial y utilizada únicamente con propósitos académicos y científicos.

Firma y sello del director(a) del centro

Universidad Nacional Autónoma de Honduras | CIUDAD UNIVERSITARIA | Tegucigalpa M.D.C. Honduras C.A. | www.unah.edu.hn

Anexo D

Plantilla propuesta por Hernández- Nieto para validar contenido de prueba de inglés como lengua extranjera (EFL) que permite evaluar la competencia lectora en estudiantes de educación primaria.

Hernández-Nieto recomienda varios aspectos del test a evaluar. Entre ellos se encuentran, los siguientes.

1. **Pertinencia:** Que tanto se corresponden el enunciado del ítem y lo que se quiere medir.
2. **Claridad conceptual:** Se examina hasta qué punto la redacción del ítem no genera confusión o contradicción con los conceptos.
3. **Redacción y terminología:** Si la sintaxis y la terminología empleada en el ítem son las adecuadas de acuerdo con el tema de estudio.
4. **Respuesta correcta:** Si el ítem ofrece una respuesta correcta que está de acuerdo con lo enunciado con el ítem.
5. **Distractores apropiados:** Los enunciados de los distractores (o respuestas incorrectas) tienen un grado de plausibilidad adecuada.
6. **Niveles de dificultad:** Los niveles de dificultad de los ítems son los apropiados y pueden tener un carácter ascendente.
7. **Formato:** La forma en cómo se presentan los ítems y sus posibles respuestas están claros.

Escala para validar cada ítem del test:

1. Inaceptable. 2. Deficiencia. 3. Regular. 4. Bueno. 5. Excelente.

Instrucciones: Coloque una X en el área de evaluación donde considere como debe ser evaluado cada ítem. Si tiene observaciones escríbalas en el espacio de observaciones. Al terminar de evaluar cada ítem coloque el total de la sumatoria de los valores escogidos.

ítem	Indicadores	Observaciones	Evaluación					Total
	Generales		1	2	3	4	5	
1	Pertinencia							
	Claridad							
	Redacción							
	Respuesta							
	Distractores							

	Dificultad							
	Formato							
Observaciones Generales:								

Nota: Adaptada de “El tema de validez de contenido en la educación y la propuesta de Hernández-Nieto”, por R. Sánchez, 2021, Latin American Journal of Physics Education, 15(3), p. 3309-5.

Anexo E

Taller de validez de contenido por jueces expertos



Fotografía tomada por: Calderon, 2024 en el Edificio 1847 de la Universidad Nacional Autónoma de Honduras, Ciudad Universitaria de Tegucigalpa Francisco Morazán. Honduras. Archivo personal

Anexo F

Tabla de Validez de contenido CVC Hernández- Nieto

Ítem	Juez	Juez	Juez	Juez	S	M	S/M	CVC	Pi	CVCi
		1	2	3	4					
1	32	35	34	33	134	35	3.83	0.96	0.00	0.95
2	35	35	35	35	140	35	4.00	1.00	0.00	1.00
3	30	32	30	30	122	35	3.49	0.87	0.00	0.87
4	35	35	34	35	139	35	3.97	0.99	0.00	0.99
5	35	35	35	35	140	35	4.00	1.00	0.00	1.00
6	35	35	35	35	140	35	4.00	1.00	0.00	1.00
7	34	35	34	33	136	35	3.89	0.97	0.00	0.97
8	35	35	35	35	140	35	4.00	1.00	0.00	1.00
9	30	30	32	30	122	35	3.49	0.87	0.00	0.87
10	35	35	35	35	140	35	4.00	1.00	0.00	1.00
11	35	35	35	35	140	35	4.00	1.00	0.00	1.00
12	30	30	28	25	113	35	3.23	0.81	0.00	0.80
13	25	25	26	25	101	35	2.89	0.72	0.00	0.72
14	25	30	28	28	111	35	3.17	0.79	0.00	0.79
15	25	25	28	30	108	35	3.09	0.77	0.00	0.77
16	35	35	35	35	140	35	4.00	1.00	0.00	1.00
17	35	35	35	35	140	35	4.00	1.00	0.00	1.00

18	30	16	35	35	116	35	3.31	0.83	0.00	0.82
19	34	30	35	30	129	35	3.69	0.92	0.00	0.92
20	35	35	35	35	140	35	4.00	1.00	0.00	1.00
21	35	35	35	35	140	35	4.00	1.00	0.00	1.00
22	35	35	35	35	140	35	4.00	1.00	0.00	1.00
23	35	35	35	35	140	35	4.00	1.00	0.00	1.00
24	28	7	30	30	95	35	2.71	0.68	0.00	0.67
25	34	35	35	34	138	35	3.94	0.99	0.00	0.98
26	35	35	35	35	140	35	4.00	1.00	0.00	1.00
27	25	20	25	25	95	35	2.71	0.68	0.00	0.67
28	35	35	35	35	140	35	4.00	1.00	0.00	1.00
29	35	35	35	35	140	35	4.00	1.00	0.00	1.00
30	35	35	35	35	140	35	4.00	1.00	0.00	1.00
31	25	26	25	20	96	35	2.74	0.69	0.00	0.68
32	35	35	35	35	140	35	4.00	1.00	0.00	1.00
33	35	35	35	35	140	35	4.00	1.00	0.00	1.00
34	35	35	35	35	140	35	4.00	1.00	0.00	1.00
35	30	35	35	35	135	35	3.86	0.96	0.00	0.96
36	35	35	35	35	140	35	4.00	1.00	0.00	1.00
37	35	35	35	35	140	35	4.00	1.00	0.00	1.00
38	35	35	35	35	140	35	4.00	1.00	0.00	1.00

39	35	10	30	35	110	35	3.14	0.79	0.00	0.78
40	34	35	34	35	138	35	3.94	0.99	0.00	0.98
41	35	35	35	35	140	35	4.00	1.00	0.00	1.00
42	30	32	34	35	131	35	3.74	0.94	0.00	0.93
43	35	35	35	35	140	35	4.00	1.00	0.00	1.00
44	35	35	34	35	139	35	3.97	0.99	0.00	0.99
45	35	35	33	27	130	35	3.71	0.93	0.00	0.92
46	30	32	35	35	132	35	3.77	0.94	0.00	0.94
47	35	35	35	35	140	35	4.00	1.00	0.00	1.00
48	35	35	35	35	140	35	4.00	1.00	0.00	1.00
49	35	35	35	35	140	35	4.00	1.00	0.00	1.00
50	35	35	35	35	140	35	4.00	1.00	0.00	1.00

Nota: Elaboración propia

Anexo G

Prueba piloto

UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS
MAESTRÍA EN PSICOMETRÍA Y EVALUACIÓN EDUCATIVA



Prueba de Fin de Grado para evaluar la competencia lectora
de inglés (EFL) para estudiantes de primer ciclo de Educación Primaria

Name: _____

score

Tipo selección única**Instruction: Read the text and circle the best option to answer each question.****Annie's Vacation**

On my last vacation my family and I travelled to Roatan. We stayed in a beautiful hotel. We went to the beach every day and swam in the ocean. My mom took a lot of photos and my dad played soccer on the beach. I was very sad to come back home.

**1. Who travel to Roatan for vacations?**

- a. Annie and her friends
- b. Annie and her family
- c. Annie and her teachers

2. Where did Annie travel on vacation?

- a. Comayagua
- b. Choluteca
- c. Roatán

Instruction: Read the text and circle the best option to answer each question.

Tommy loves soccer

Tommy loves playing soccer. He and his friends practice every day after school. He kicks the ball and learns new skills. Tommy's favorite part is scoring goals. He feels proud when he helps his team win.

3. What sport does Tommy love to play?

- a. basketball
- b. soccer
- c. tennis

4. What is Tommy's favorite part of soccer?

- a. running with the ball
- b. kicking the ball
- c. scoring goals



Instruction: Read the text and circle the best option to complete each statement.

“My best friend Hashiko”

My friend Roberto likes to watch movies every weekend. His favorite movie is “Hashiko” because is about a charming dog which loves its owner and waits for him at the train station every day when he comes back home.



5. My friend Roberto likes to watch _____ every weekend.

- a. Songs
- b. shows
- c. Movies

6. "Hashiko" is a _____ in the story.

- a. turtle
- b. cat
- c. dog

Instruction: Read the following text and answer the questions.

The Weather Forecast in Honduras

Welcome to the Honduras Weather forecast. In the north of the country is windy, there is some rain too, so do not leave your home without an umbrella. The temperature is around 30° centigrade. In the east its rainy all day today. In the west and middle of the country there may be a thunderstorm in the afternoon, the temperature is at 29 °centigrade. The South of the country has the best weather today. Its sunny this afternoon. The temperature is about 27° centigrade.

7. According to the report how is the weather in the north of Honduras?

- a. sunny
- b. windy
- c. cloudy

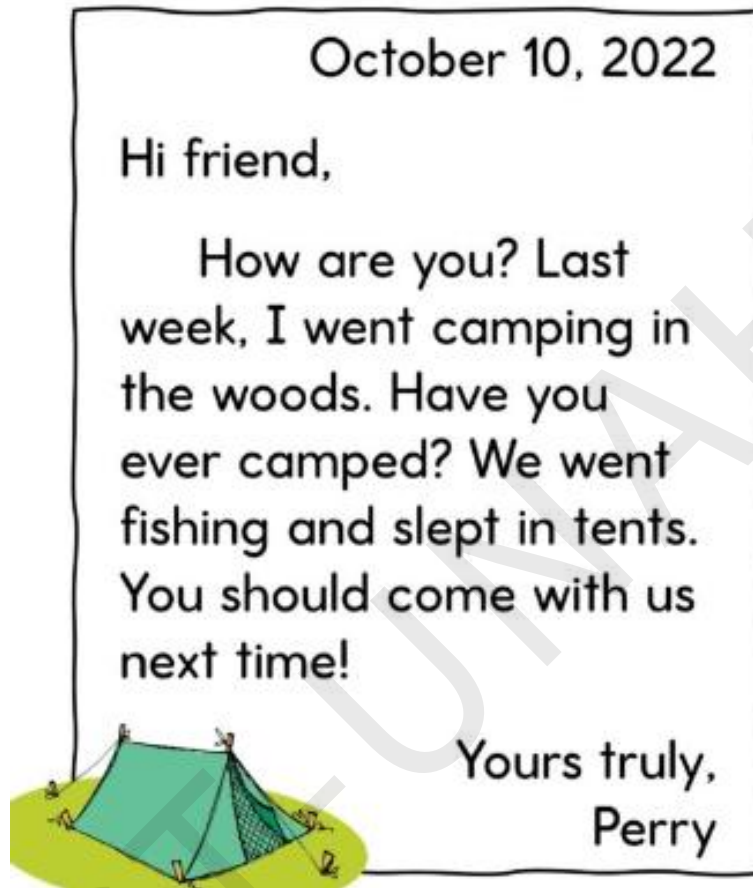
8. Which place of Honduras has the best weather today?

- a. north
- b. west
- c. south

9. What is the temperature in the south of Honduras?

- a. 29° degrees
- b. 27° degrees
- c. 30° degrees

Instruction: Look at the picture. Read the Friendly letter. Choose the correct option to each question.



According to the friendly letter answer the following questions

10. What is the letter about?

- a. camping in the woods
- b. camping in the lake
- c. camping in the park

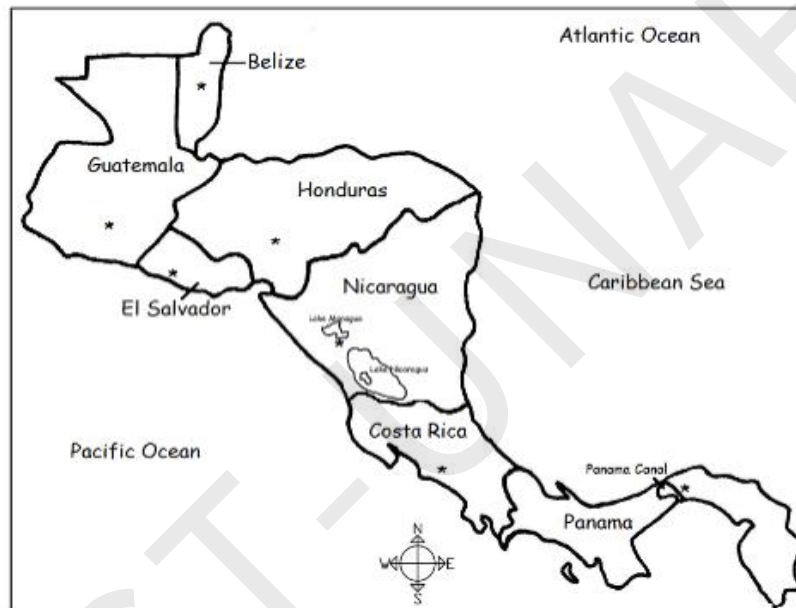
11. Which activities had in the camping?

- a. fishing and playing soccer
- b. fishing and slept in tents
- c. fishing and running by the woods

Instruction: Read the following text and answer the questions.

Central America

Central America consist of seven countries that connect North America to South America, with Mexico to the north and Colombia to the southeast. The seven countries of Central América are: Belize, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, and Panamá. The main language of Central America is Spanish, although English is the official language of Belize.



According to the text “Central America” circle the correct option.

12. Central America consists in seven_____.

- a. countries b. oceans c. maps

13. The main language of Central America is _____

- a. Portuguese b. Spanish c. English

14. According to the text, which countries belong to Central America.

- a. Belize, Guatemala, Colombia, Honduras, Nicaragua, Costa Rica and Panamá
 b. Belize, Guatemala, México, Honduras, Nicaragua, Costa Rica and Panamá
 c. Belize, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, and Panamá

Instruction: Read the text about “Christmas” and answer the questions.

Christmas

Story by Have Fun Teaching

Christmas is in December. It is a season with many traditions. People think about family, friends, and giving. Most families put up a tree, hang decorations and give presents. Christmas Eve is December 24th. It is the night before Christmas. It is a great time to sing carols. Will you remember to put out milk and cookies for Santa? Christmas Day is December 25th. There is going to be a lot of food to eat that day. It is a time to spend with family and people you love.



According to the text “Christmas” answer the following.

15. Christmas is in ... a. December b. January c. November

Choose the picture to complete the sentence. Write an X on the correct choice.

16. Christmas is a season that people think about...

a.



b.



c.



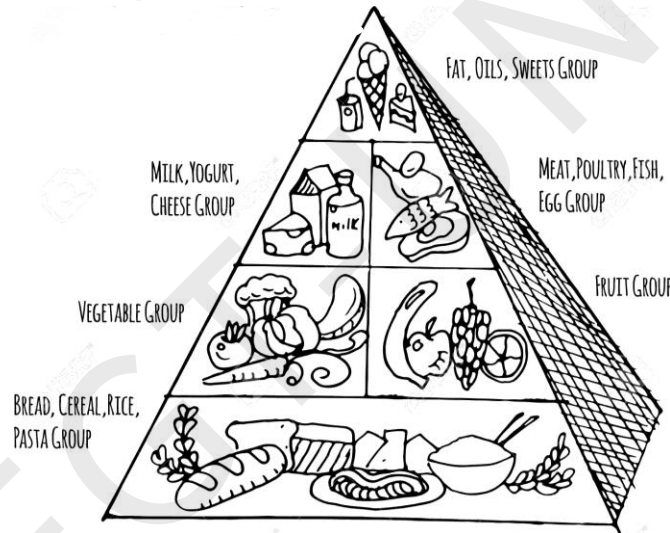
17. According to the text “Christmas” what is Christmas Eve?

- a. It is the night before Christmas
- b. It is the morning after Christmas
- c. It is the night after Christmas

Instruction: Look at the picture. Read the text and choose the correct answer to the question.

Healthy Food

Healthy food is tasty and good for you. You need to grow and feel good. You can find **grains** in bread and rice. Milk, cheese and yogurt are also important. These will give you strong bones. Meat, eggs, fish, and nuts give you **protein**. **Fruits** and **vegetables** are good to eat. Eat as many as you can every day. Chips and cookies are tasty, but **eat only some times**.



According to the text “Healthy Food” choose the correct option.

18. What can you find in bread and rice?

- a. fruits
- b. grains
- c. milk

19. Write three food you can eat so your bones will be strong.

- a. candies, sugar, salt
- b. meat, cheese, fruits
- c. milk, cheese, yogurt

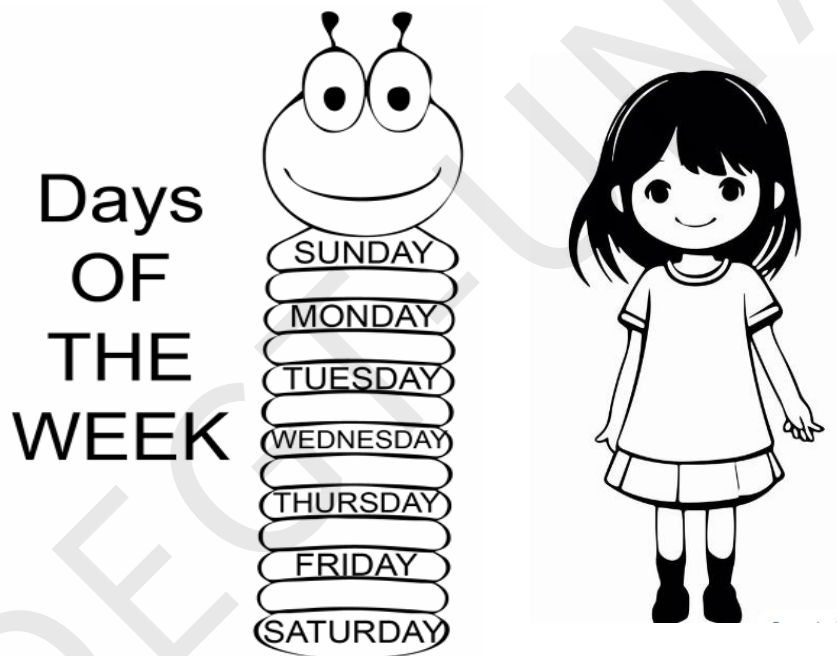
20. Meat, eggs, fish, and nuts give you...

- a. protein
- b. vitamins
- c. grains

Instruction: Look at the picture. Read the text and choose the correct answer to the question

Seven Days of the week

This is Mary. She is ten years old. Every Monday, Mary reads Chinese books with her friends. Every Tuesday, she rides a bike to school. On Wednesdays, she plays music at school. On Thursdays, she sings English songs with her friends. Every Friday, she cooks eggs at home. Saturday and Sunday, she rests.



According to the text “Seven Days of the Week” Circle the correct option.

- | | | | |
|--|-----|----|----------------|
| 21. Every Monday she reads Chinese books | yes | no | it doesn't say |
| 22. Mary rides a bike to school on Fridays | yes | no | it doesn't say |
| 23. On Wednesday she plays music at home | yes | no | it doesn't say |
| 24. Every Friday she cooks eggs | yes | no | it doesn't say |

Instruction: Read the text about “Valentine’s Day “and answer the questions.

Valentine’s Day

Valentine’s Day is always February 14th. It is a holiday about love. Hearts and cupids are symbol of Valentine’s Day. There are many legends about the original Valentine that the holiday’s name come from. These days, Valentine’s Day is about love and appreciation. You can show your family how much you love them.



Using the information about “Valentine’s Day” Complete the sentence choosing the correct word.

25. Valentine’s Day is always on ...

a. February 4th

b. February 14th

February 11th

26. Valentine’s Day is a holiday about



Instruction: Read the text about “Mother’s Day “and answer the questions.

Mother’s Day

story by Andrew Frinkle

Mother’s Day is the 2nd Sunday in May. It is a day to celebrate and all that they do for us. Today, over 150 countries celebrate a Mother’s Day, but they might have a different day for it. It is important to look at what mothers bring to their families. Every mother is different. Does your mother make you healthy meals? Does she help with your homework? Does she read you bedtime stories? Does she hug you? Thank her for all she does! Mother’s Day is also a great time to visit or call your grandmothers. They’re mothers too!



Using the information about “Mother’s Day” choose the correct answer.

27. Mother’s Day is in the...

- a. 2nd Sunday in March
- b. 2nd Sunday in May
- c. 2nd Sunday in June

28. Over _____ countries celebrate Mother’s Day.

- a. 50
- b. 115
- c. 150

29. Grandmothers are mothers too.

- a. yes
- b. no
- c. I don’t know

Instruction: Read the text about “Angela loves school “and answer correctly.

Angela loves school

Angela is an eight years old student. She is very smart. She loves to learn about countries. Today she learned about Mexico. Angela wants to visit Mexico because her grandparents live there. When she is on vacation from school, she plans to travel there. She wants to visit museums, theaters, and many places in Mexico.



Using the information about “Angela loves school” choose the correct answer.

30. What is the girl’s name?

- a. Jessica b. Becky c. Angela

31. What does she love to learn about?

- a. countries b. animals c. sports

32. What country did she learn about today?

- a. Canadá b. Italia c. México

33. Who lives in México?

- a. Dad b. uncle c. grandparents

Instruction: Read the text about “My best friend “Identify the main idea.

My Best Friend

Dogs are often called man's best friend, but they also help humans in many ways. Some dogs hunt or retrieve. Some dogs guide the blind people. Others guard animals and property. They are good pets. They move its tail when they are happy. Dogs are good friends.



34. Using the information about “My best friend” choose the main idea of the text.

- a. Dogs help young people and other dogs.
- b. Dogs are often called man's best friend.
- c. Some dogs hunt and helps blind people.

35. An important detail that support the main idea is ...

- a. Some dogs guide blind people and guard property.
- b. Dogs move its tail when they are happy.
- c. Dogs eat a lot of food when they are sad.

Instruction: Read the text about “Going to the Market “and answer correctly.

Going to the Market

Every Saturday Morning, Josie went to market with her mother. She saw many kinds of fish, vegetables, and fruits. My mom bought tomatoes, mangoes, watermelon, and pineapples. Mom says that people sell vegetables they grow. They also sell flowers and fruits, some of them sell bread. There are many people in the market. They are very friendly and nice.

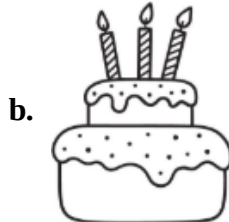


Using the information about “Going to the market” circle the correct option

- | | | |
|---------------------------------------|--------------|-------------------|
| 36. There are many people in a market | a. Yes b. no | c. It doesn't say |
| 37. People sell vegetables they grow | a. yes b. no | c. It doesn't say |
| 38. People in the market are angry | a. yes b. no | c. it doesn't say |

Using the information about “Going to the market” write an X in the correct option.

- 39. You can buy in the market ...**



Nota: Fuente Elaboracion propia

Anexo H



Fotografía tomada por: Calderon, 2024 en la Escuela 4 de Julio del barrio La Esperanza del municipio del Distrito Central, Francisco Morazán. Archivo personal

Anexo I



Fotografía tomada por: Calderon, 2024 en el C. E. B. Juan Ramon Molina
del Municipio del Distrito Central.
Francisco Morazán. Archivo personal

Anexo J

Parámetros de discriminación (α) y dificultad (β) estimados para los 39 ítems del instrumento de evaluación de la competencia lectora en inglés (EFL), según el modelo 2PL.

Ítem	Discriminación (α)	Dificultad (β)
It_1	0.173	-3.777
It_2	-0.132	2.693
It_3	-0.087	-0.080
It_4	0.445	-0.877
It_5	-0.909	-0.952
It_6	0.780	-1.205
It_7	-2.200	0.350
It_8	0.017	-107.271
It_9	0.057	29.998
It_10	-0.091	21.068
It_11	0.469	0.546
It_12	2.445	-0.190
It_13	0.465	-1.352
It_14	0.051	-5.625
It_15	0.156	-8.262
It_16	3.440	-0.619
It_17	-0.007	-0.080
It_18	63.434	-0.084
It_19	-0.140	-6.484
It_20	0.097	9.800
It_21	0.144	-3.599
It_22	1.572	-1.582
It_23	-0.012	139.498
It_24	0.006	-230.938
It_25	-0.794	0.229
It_26	0.208	-7.671
It_27	0.521	-1.378

It_28	-0.015	65.272
It_29	-0.583	-0.165
It_30	-0.602	5.639
It_31	-0.141	10.228
It_32	0.012	-182.022
It_33	0.752	-1.017
It_34	-0.594	-2.474
It_35	0.311	3.243
It_36	0.034	-36.508
It_37	0.182	-7.077
It_38	0.213	-5.796
It_39	-0.340	5.222

Nota: Elaboracion Propia. Estimaciones realizadas con base en el modelo logístico de dos parámetros (2PL) aplicado a los ítems del instrumento.