# Análisis de ítems para prueba de clasificación en inglés en una Universidad colombiana[1]

# Item analysis for an English placement test at a Colombian university

DOI: http://dx.doi.org/10.17981/culteduso c.11.2.2020.11

Recibido: 13 de abril de 2020 Aceptado: 19 de junio de 2020 Publicado: 09 de julio de 2020

**Alexander Ramirez Espinosa**
Universidad del Valle. Cali (Colombia)
alexander.ramirez.e@correounivalle.edu.co

Para citar este artículo:
Ramirez, A. (2020). Análisis de ítems para prueba de clasificación en inglés en una Universidad colombiana. *Cultura, Educación y Sociedad*, 11(2). 177-190. DOI: http://dx.doi.org/10.17981/culteduso c.11.2.2020.11

*Resumen*

Este artículo reporta las mediciones del índice de dificultad (IF) y del índice de discriminación (ID) en una prueba de clasificación en inglés diseñada por y para una Universidad pública colombiana. La prueba de clasificación estuvo compuesta de una sección de gramática y vocabulario, una sección de comprensión oral y una sección de comprensión escrita. La medición de los índices permitió categorizar ítems en los siguientes niveles: excelentes, aceptables e incorrectos para su posterior corrección o descarte. Los resultados del estudio sugieren que los ítems hallados incorrectos corresponden a preguntas cuyas instrucciones son ambiguas o cuyas opciones de respuesta contenían distractores poco funcionales; entre las conclusiones más importantes del estudio se confirma la falta de literacidad en evaluación de lenguas (LAL, por sus siglas en inglés) entre los profesores de lengua extranjera y se reafirma la necesidad de un entrenamiento explícito en diseño de ítems.

*Palabras clave:* Evaluación de lenguas extranjeras; diseño de exámenes; examen de clasificación; análisis de ítems; literacidad en evaluación de lenguas

*Abstract*

This article reports the measurements of difficulty index (IF) and discrimination index (ID) in an English placement test designed by and for a Colombian public university. The test was made up of a grammar and vocabulary section, an oral comprehension section and a written comprehension section. The measurement of the indexes allowed categorizing excellent, acceptable and defective items for their subsequent correction or discard. The results of the study suggest that the items found to be defective correspond to questions that portrayed either ambiguous instructions or non-functional distractors; among the main conclusions the study confirms the lack of language assessment literacy (LAL) in foreign language teachers and endorses the need for explicit training in item design.

*Keywords:* Foreign languages evaluation; testing; placement test; item analysis; language testing literacy

[1] Este artículo es producto del proyecto de investigación "Pilotaje de una Prueba de Clasificación de Inglés para Estudiantes de Primer Semestre en la Universidad del Valle: Análisis de Ítems", que se llevó a cabo en la Escuela de Ciencias del Lenguaje de la Universidad del Valle. Este proyecto fue auspiciado por la Vicerrectoría de Investigaciones de la Facultad de Humanidades, bajo el código CI-4383.

## Introducción

La evaluación es una tarea inherente al oficio de los docentes, y en el campo específico de la enseñanza de lenguas extranjeras, la evaluación formal o informal del nivel de adquisición de una lengua extranjera por parte de los aprendices es una actividad constante en el quehacer del maestro. Sin embargo, la literatura académica sobre evaluación de lenguas evidencia que, en el ámbito colombiano, los profesores de lengua no están lo suficientemente formados en el tema de evaluación ni en el diseño de pruebas confiables (López & Bernal, 2009; Giraldo 2018a, 2018b). Aun así, el docente de lenguas extranjeras debe diseñar, adaptar y aplicar pruebas constantemente, y una vez obtenidos los resultados numéricos, se espera que este docente sea capaz de hacer conversiones y equivalencias a diferentes métricas, en un afán por medir el nivel de dominio de una lengua, o en este caso del inglés como lengua extranjera, de los estudiantes con estándares comparables internacionalmente.

A esta carencia de una sólida formación en evaluación, se suma la falta de investigación y estudios empíricos sobre el diseño de pruebas de lengua en Colombia: pese a la gran cantidad de pruebas que se administran y se exigen diariamente de manera oficial o institucional con propósitos diversos, una mirada a las revistas académicas colombianas revela que no hay suficiente producción de conocimiento al respecto.

La evaluación mediante el uso de las pruebas estandarizadas, sin embargo, es una actividad que subyace a la profesión docente y a las prácticas pedagógicas de diversas instituciones; sin embargo, si se determina la existencia de carencias en formación y en investigación sobre prácticas evaluativas, es posible que tanto los docentes como las instituciones, en general, estén cometiendo errores e irregularidades en sus procesos de evaluación las cuales impactan todo el proceso pedagógico, toda vez que la falta de rigor en el diseño de las pruebas "puede desorientar la enseñanza y el aprendizaje de los idiomas" (Giraldo, 2019, p. 124)[2], así como las decisiones que se derivan de sus resultados.

Teniendo en cuenta lo anterior, y ante una necesidad institucional particular en la Universidad del Valle, en Cali-Colombia, se diseñó un proyecto de investigación con un doble propósito: por un lado, diseñar una prueba de clasificación en inglés con altos estándares de calidad; y por otro lado capacitar a un grupo de docentes para iniciar una línea de estudio en diseño de pruebas y sus posteriores análisis estadísticos. El proyecto produjo un test compuesto de 52 ítems de selección múltiple, repartidos en tres secciones que corresponden a la comprensión de lectura, la comprensión oral y la evaluación de estructuras formales de la lengua (gramática).

El pilotaje de la prueba comprendió la medición y el análisis de los índices de discriminación y dificultad de los 52 ítems; estos índices ayudan a determinar la validez y confiabilidad de la prueba y permiten hacer correcciones que se traducen en su constante mejoramiento (Carr, 2011). Así las cosas, este artículo presenta los resultados de dichas mediciones, así como las reflexiones sobre diversos retos académicos y administrativos experimentados en el proceso, con el fin de proveer una mirada general de lo que implica la ardua empresa de diseñar pruebas y estudiarlas de manera estadística.

[2] Todas las traducciones de referencias originalmente escritas en inglés son realizadas por el autor.

## Referentes Conceptuales

### *La Necesidad de Diseñar Pruebas Propias*

El diseño, la aplicación y la posterior investigación sobre pruebas propias de la institución son aspectos deseables dentro de una institución, no sólo porque puede ayudar a minimizar la compra de pruebas comerciales -que además de implicar altos costos, no están diseñadas para las necesidades de una institución o contexto particular- sino porque también empodera a los maestros y contribuye a su formación continua en términos de Literacidad en la Evaluación de Lenguas (LAL). Esto último se traduce en una mayor experiencia y diversificación de técnicas de evaluación, así como en el desarrollo de la reflexión constante y el pensamiento crítico hacia los procesos evaluativos.

A este respecto, Tomlinson (2005) sostiene que "si bien es obviamente importante que las pruebas sean justas, válidas y confiables, lo más importante de todo es que las pruebas proporcionen oportunidades útiles para el aprendizaje" (p. 40); se trata de oportunidades de aprendizaje no sólo para quienes toman las pruebas, sino también para aquellos que las diseñan y las aplican: "El objetivo principal de las pruebas de lengua es proporcionar oportunidades de aprendizaje, tanto para los estudiantes que están siendo evaluados como para los profesionales que administran las pruebas" (Tomlinson, 2005, p. 39). En esa misma línea, Taylor y Geranpayeh (2011) aseguran que el diseño de una nueva prueba tiene "un potencial intrínseco como instrumento de investigación cuyos resultados ayudarán a enriquecer nuestra comprensión de la naturaleza del dominio del idioma para que podamos desarrollar mejores pruebas en el futuro" (p. 94); a esto se suma la posibilidad de aprender y profundizar sobre la construcción de instrumentos de evaluación, lo cual se traduce en opciones de desarrollo profesional a corto, mediano y largo plazo para los docentes de lenguas extranjeras.

Todo lo anterior, en consonancia con los postulados de López y Bernal (2009) y Giraldo (2018a), argumentan a favor de la imperiosa necesidad de capacitar a los profesores de idiomas en el diseño, análisis e investigación de exámenes propios. Más allá de ser simplemente conscientes de los diferentes tipos de pruebas y sus usos, los profesores de idiomas requieren "conocimiento sobre cómo escribir, administrar y analizar pruebas" (Inbar-Lourie, 2013, p. 32), de manera que dichos profesores pasen de ser usuarios y consumidores de exámenes, a ser diseñadores, productores y críticos de la evaluación. En este sentido, vale la pena recordar lo que Giraldo (2018a) menciona, haciendo eco de López y Bernal (2009), quienes indican que los docentes de lengua que han recibido capacitación formal en evaluación, la utilizan para mejorar los procesos de enseñanza y el aprendizaje, mientras que aquellos que no han recibido ningún entrenamiento formal, sólo la usan como un mecanismo de obtener calificaciones numéricas.

### *Las Pruebas de Clasificación en Lengua Extranjera*

Fernández (2007) define las pruebas de clasificación como aquellas que "pueden servir para estimar el nivel lingüístico de un grupo de alumnos y, al mismo tiempo, ubicarlos en distintos grupos" (p. 189). Son pruebas que se relacionan estrechamente con programas

de estudios en particular y con su currículo (Brown, 2011). Estas pruebas resultan muy útiles en contextos universitarios, justo después de la admisión de estudiantes, sobre todo porque previo al ingreso es difícil obtener registros confiables del nivel de conocimiento de una lengua de los admitidos (en este caso, inglés), o porque estos proporcionan diversos certificados y resultados de pruebas que no son comunes, con escalas de calificación difíciles de interpretar, con evidencias incompletas, etc. (Wall, Claphman & Aldersen, 1994).

Estas pruebas se caracterizan por ser cortas, pues casi siempre es necesario obtener los resultados lo más pronto posible para tomar decisiones administrativas y logísticas (Fernández, 2007). Su estructura interna generalmente incluye una sección de gramática, una de comprensión de lectura y una de comprensión oral (ver, por ejemplo, las pruebas comerciales OOPT y QPT). Las habilidades de producción oral y escrita suelen desaparecer, en virtud de la premura del tiempo para obtener resultados y de la tardanza que implicaría la corrección; sin embargo, sí hay reportes de pruebas de clasificación las cuales incluyen entrevistas y sección de escritura (Paltridge, 1992), sobre todo cuando la población de examinados es pequeña.

### *Algunas Características de las Pruebas de Selección Múltiple*

Las preguntas de selección múltiple son un tipo de ítem que se componen de una raíz o enunciado (*stem*) -que contiene la instrucción- y tres o cuatro opciones de respuesta, entre las que se encuentra la opción correcta (*key*). Palés-Argullós (2010) las caracteriza como "pruebas objetivas por su demostrada alta fiabilidad, aunque para ello deben estar bien diseñadas" (p.149). Y es que el diseño adecuado de este tipo de ítems requiere un cierto nivel de experiencia, sin embargo, gracias a su ubicuidad, se suele pensar que son ítems fáciles de diseñar; al respecto Carr (2011) asegura que lo fácil es diseñarlos de manera incorrecta.

Dos de los mayores retos en el diseño de estos ítems consiste en escribir enunciados que no den lugar a la ambigüedad, y en escoger distractores de buena calidad. Sobre este último punto, cabe mencionar que un buen distractor es aquel que es susceptible de ser escogido por algunos de los examinados (Carr, 2011). Un distractor que no es atractivo a los ojos de los examinados como posible opción es un distractor disfuncional, que no sirve para medir conocimientos y no aporta nada positivo al ítem ni al examen, sino que por el contrario, ejerce un impacto negativo en los examinados (Malau-Aduli & Zimitat, 2012). En este sentido, el principio para el diseño de distractores es que cada uno de ellos "debe ser plausibles, sin llegar a estar tan cerca de ser correctos que termine habiendo dos respuestas correctas" (Carr, 2011, p. 92).

Toksöz y Ertunç (2017) afirman que los tests de selección múltiple deben su amplia aceptación en parte, al hecho de que funcionan como "guardianes" (*gatekeepers*) de manera más efectiva que los tests con otro tipo de preguntas. A esto se suma la familiaridad que traen los estudiantes de su experiencia en la secundaria y el criterio de practicidad en términos de corrección. Tal como lo plantean Toksöz y Ertunç (2017) para el contexto turco, en Colombia el uso de ítems de selección múltiple es bastante popular en diferentes niveles de instrucción, sin embargo, en el contexto colombiano no hay suficientes estudios empíricos que analicen este tipo de ítems en términos de su índice de dificultad, índice de discriminación o eficiencia de los distractores.

## *El Análisis de Ítems en Pruebas de Lengua*

Es necesario abrir este apartado señalando que los análisis estadísticos de pruebas para conocer el nivel de dominio de una lengua extranjera suelen concebirse como una tarea propia de instituciones comerciales -y sus investigadores- que aplican evaluaciones a gran escala; de ser así, dicha tarea estaría fuera del dominio del docente de lenguas y de su contexto inmediato en el aula de clases. Sin embargo, a este respecto Carr (2011) advierte que los estudios de estadísticas descriptivas en pruebas de lengua son necesarios tanto en contextos de evaluación masiva como en ámbitos más reducidos, poniendo como argumento las siguientes cinco razones:

- Las estadísticas descriptivas pueden ayudarnos a dilucidar si un examen es apropiado o no para el propósito con el que se va a aplicar y para la población que se evaluará.
- Los estudios estadísticos de una prueba permiten establecer coeficientes de correlación, entendidos como la posibilidad de establecer comparaciones entre individuos, grupos, escalas, etc.
- Las fórmulas utilizadas en las estadísticas descriptivas ayudan a calcular la confiabilidad y la validez de una prueba
- Las estadísticas permiten visualizar, de manera sintética, el desempeño individual de un examinado y el desempeño colectivo de uno o más grupos, lo cual redunda en la posibilidad de comparaciones y correlaciones que pueden ser muy útiles a nivel curricular
- Las estadísticas permiten comunicar información sobre las pruebas de manera precisa y significativa, tal como lo establece el Código de Ética de la Asociación Internacional de pruebas de Lengua (ILTA, por sus siglas en inglés) (ILTA, 2001).

Con todo lo anterior, y ante la necesidad de consolidar una prueba de clasificación válida y confiable, se optó iniciar los estudios estadísticos de la misma con la medición de los Índices de Dificultad (en adelante IF, por sus siglas en inglés) y de discriminación (en adelante ID, por sus siglas en inglés). Si bien estos no son los únicos estudios estadísticos que se pueden aplicar a una prueba, sí constituyen un punto de partida, según lo establecen los antecedentes consultados en la revisión bibliográfica (Janssen & Meier, 2013; Toksöz & Ertunç, 2017; Sugianto, 2020). A continuación, se definen ambos índices.

### *El Índice de Dificultad (IF) y el Índice de Discriminación (ID)*

Estos dos términos provienen de la teoría clásica de evaluación y para los propósitos del presente estudio se tomaron las definiciones en las que coinciden todos los teóricos expertos en el ámbito particular de la evaluación de lenguas (Brown, 2004; Bachman, 2004; Carr, 2011). Así, el IF se define como el grado de complejidad que tiene un ítem para un grupo determinado de examinados. En otras palabras, la medición de este índice en cada ítem del examen permite determinar si las preguntas diseñadas son apropiadas, demasiado fáciles, difíciles o incluso imposibles de resolver para la población que se pretende evaluar.

Si un ítem es demasiado fácil al punto que todos los examinados de nivel más bajo lo pueden resolver, o si por el contrario es tan difícil que ninguno de los examinados de nivel más alto logra resolverlo satisfactoriamente, se podría sospechar que hay defectos por corregir en el diseño de ese ítem en particular. Cabe aclarar que la medición de este índice no constituye el único argumento para descartar un ítem, sino que es necesario completar el análisis con la medición del ID (Carr, 2011). Siguiendo a Toksöz y Ertunç (2017) la fórmula para calcular este índice puede resumirse así:

$$IF = \frac{\text{Cantidad de examinados que resolvieron el ítem correctamente}}{\text{Total de examinados que respondieron el ítem}}$$

Por otra parte, el ID hace referencia al potencial que tiene un ítem para distinguir a los examinados con mejor nivel de los examinados con nivel inferior (Brown, 2004; Carr, 2011; Toksöz & Ertunç, 2017). Si un ítem obtiene respuestas correctas de la mayoría de examinados de nivel alto, y respuestas incorrectas de la mayoría de examinados de nivel bajo, entonces dicho ítem es apropiado para diferenciar entre examinados de proficiencia alta y de proficiencia baja. Si, por el contrario, un ítem obtuviera respuestas correctas de la mayoría de examinados con proficiencia baja y, además, obtuviera respuestas incorrectas por parte de los examinados de proficiencia alta, dicho ítem es defectuoso y debe considerarse su revisión o incluso su descarte. Cabe recordar, de acuerdo a Brown (2004), los ítems con mejor rendimiento discriminatorio tienen un valor cercano o igual a 1.0, mientras que los ítems con baja discriminación tienen un valor más cercano a cero.

## Metodología

### *Test y Participantes*

Para este estudio se aplicó la prueba UVEPLAT a 815 estudiantes admitidos a primer semestre de diversos programas de la Universidad del Valle, en Cali-Colombia. La población de admitidos estuvo compuesta por un 61% de mujeres y 49% de hombres, cuya edad promedio fue 17 años, todos ellos provenientes en su mayoría de instituciones de educación pública (82%). La prueba UVEPLAT (Univalle's English Placement Test) comprende 52 ítems diferentes y clasifica a los examinados en los niveles A1, A2, B1 y B2 de acuerdo con el Marco Común Europeo de Referencia para las Lenguas (MCER). Estos cuatro niveles se corresponden con los cuatro cursos obligatorios que deben inscribir los estudiantes admitidos a la Universidad.

La prueba UVEPLAT fue diseñada por docentes de la Escuela de Ciencias del Lenguaje, quienes emprendieron un proceso de capacitación en el diseño de pruebas y lo complementaron con su experiencia como docentes de inglés con propósitos generales y académicos, en los cuatro niveles obligatorios que establece la institución. La prueba requiere una hora para su finalización y evalúa a los examinados en comprensión de lectura, comprensión auditiva, estructuras gramaticales y vocabulario. Los 52 ítems que componen la prueba corresponden a preguntas de selección múltiple, estructura escogida por la familiaridad que los estudiantes tienen con este tipo de preguntas desde su vida escolar previa a la Universidad; a este respecto seresalta la importancia de que las pruebas reflejen las maneras en las que el examinado aprende y usa la lengua en los contextos escolares previos. La

prueba se encuentra alojada en la Plataforma Moodle de la Universidad (Campus Virtual) lo cual permitió su aplicación en formato electrónico y su calificación de manera automática e inmediata.

### *Recolección de datos*

Los datos recogidos corresponden a las 52 respuestas que cada estudiante proporcionó en su prueba (42.380 datos analizados en total). Se excluyeron aquellos examinados que no completaron la prueba porque excedieron el tiempo de la plataforma, o porque decidieron abandonar la prueba en algún momento. Los datos se recogieron durante el segundo semestre de 2018, entre el 27 de agosto y el 15 de septiembre. Gracias al hecho de que la prueba está alojada en la plataforma Moodle, es sencillo visualizar los datos y trasladarlos a hojas de cálculo para su posterior análisis.

### *Análisis de datos*

Los datos fueron procesados a través del aplicativo estadístico que proporciona la plataforma de Moodle; además, se corroboraron las mediciones de Moodle para los índices IF e ID haciendo el ingreso manual de datos en plantillas de trabajo proporcionadas por Carr (2011). Para la interpretación de los resultados obtenidos se han adaptado las escalas clásicas propuestas por Ebel y Frisbie (1986), reconocidas en el ámbito de la evaluación en lenguas por autores como Brown (2004). La Tabla 1 muestra la escala para la interpretación del IF y la Tabla 2 muestra la escala para la interpretación del ID:

Tabla 1.
*Escala de Interpretación del Índice de Dificultad (IF)*

| Rangos | Valoración |
|---|---|
| 0.20 - menos | Muy difícil (considerar revisión o descartar) |
| 0.21 - 0.40 | Difícil |
| 0.41 - 0.60 | Moderado |
| 0.61 - 0.80 | Fácil |
| 0.81 - más | Muy fácil (considerar revisión o descartar) |

Fuente: Autores.

Tabla 2.
*Escala de Interpretación del Índice de Discriminación (ID)*

| Rangos | Valoración |
|---|---|
| 0.40 - más | Ítem Excelente |
| 0.30 - 0.39 | Ítem Bueno |
| 0.20 - 0.29 | Ítem Aceptable |
| 0.09 - 0.19 | Ítem Deficiente |

Fuente: Autores.

Si bien los estudios sobre índices de dificultad apuntan a tener un banco de ítems que sean considerados en su mayoría moderados, cabe aclarar que para el caso la prueba UVEPLAT, cuyo propósito es clasificar a los examinados en cuatro grupos, será necesario tener un banco de ítems que vayan desde lo fácil hasta lo difícil.

## Resultados

### *Medición del Índice de Dificultad (IF)*

La Tabla 3 recoge todos los valores obtenidos del índice de dificultad obtenidos para cada ítem que compone la prueba UVEPLAT. Esta medición mostró que tres ítems dentro de la prueba (los cuales representan un 6%) estaban en el rango de ***muy fácil*** (se han marcado en celdas negras), y que 6 ítems (12%) se encuentran en el rango de ***muy difícil***, (se han marcado en celdas grises). Los 43 ítems restantes se encuentran repartidos así: entre la categoría ***difícil*** 21 ítems (41%), en la categoría ***moderado*** 16 ítems (31%), y en la categoría ***fácil*** 6 ítems (11%). Esto quiere decir que es una prueba con una complejidad considerable bien distribuida entre los ítems, lo cual es ideal en una prueba de clasificación.

Tabla 3.
*Valores de IF en la prueba UVEPLAT*

| Ítem | IF | Ítem | IF | Ítem | IF | Ítem | IF |
|---|---|---|---|---|---|---|---|
| Ítem #1 | 0.87 | Ítem #14 | 0.77 | Ítem #27 | 0.36 | Ítem #40 | 0.46 |
| Ítem #2 | 0.94 | Ítem #15 | 0.67 | Ítem #28 | 0.40 | Ítem #41 | 0.82 |
| Ítem #3 | 0.35 | Ítem #16 | 0.50 | Ítem #29 | 0.21 | Ítem #42 | 0.60 |
| Ítem #4 | 0.20 | Ítem #17 | 0.58 | Ítem #30 | 0.46 | Ítem #43 | 0.58 |
| Ítem #5 | 0.23 | Ítem #18 | 0.75 | Ítem #31 | 0.30 | Ítem #44 | 0.32 |
| Ítem #6 | 0.54 | Ítem #19 | 0.32 | Ítem #32 | 0.01 | Ítem #45 | 0.29 |
| Ítem #7 | 0.04 | Ítem #20 | 0.36 | Ítem #33 | 0.89 | Ítem #46 | 0.38 |
| Ítem #8 | 0.59 | Ítem #21 | 0.35 | Ítem #34 | 0.50 | Ítem #47 | 0.67 |
| Ítem #9 | 0.33 | Ítem #22 | 0.66 | Ítem #35 | 0.31 | Ítem #48 | 0.38 |
| Ítem #10 | 0.29 | Ítem #23 | 0.41 | Ítem #36 | 0.17 | Ítem #49 | 0.41 |
| Ítem #11 | 0.60 | Ítem #24 | 0.57 | Ítem #37 | 0.17 | Ítem #50 | 0.52 |
| Ítem #12 | 0.36 | Ítem #25 | 0.11 | Ítem #38 | 0.27 | Ítem #51 | 0.38 |
| Ítem #13 | 0.39 | Ítem #26 | 0.65 | Ítem #39 | 0.43 | Ítem #52 | 0.40 |

Fuente: Autores.

### *Medición del Índice de Discriminación (ID)*

Esta medición permitió constatar que, en general, el 81% de los ítems en la prueba UVEPLAT logran su objetivo de discriminar y clasificar a los examinados. Este 81% de ítems se encuentra repartido en las categorías de discriminación, a saber: excelente, buena y moderada, como se muestra en la Tabla 4. El 19% restante de los ítems (marcados

en celdas negras) presentan valores por debajo de 0.20, lo cual indica que son ítems que no logran diferenciar entre los estudiantes de mejor nivel y aquellos de nivel más bajo. Incluso, se puede observar que el ítem #25 presenta un valor negativo de -010; de acuerdo con la literatura, un valor negativo en la medición del ID de un ítem indica que incluso los estudiantes de más alto nivel no lograron responderlo de manera adecuada.

Tabla 4.
*Valores de ID en la prueba UVEPLAT*

| Ítem | ID | Ítem | ID | Ítem | ID | Ítem | ID |
|---|---|---|---|---|---|---|---|
| Ítem #1 | 0.23 | Ítem #14 | 0.48 | Ítem #27 | 0.21 | Ítem #40 | 0.24 |
| Ítem #2 | 0.15 | Ítem #15 | 0.48 | Ítem #28 | 0.41 | Ítem #41 | 0.31 |
| Ítem #3 | 0.38 | Ítem #16 | 0.23 | Ítem #29 | 0.14 | Ítem #42 | 0.38 |
| Ítem #4 | 0.49 | Ítem #17 | 0.33 | Ítem #30 | 0.38 | Ítem #43 | 0.58 |
| Ítem #5 | 0.04 | Ítem #18 | 0.32 | Ítem #31 | 0.01 | Ítem #44 | 0.32 |
| Ítem #6 | 0.23 | Ítem #19 | 0.20 | Ítem #32 | 0.12 | Ítem #45 | 0.29 |
| Ítem #7 | 0.47 | Ítem #20 | 0.26 | Ítem #33 | 0.22 | Ítem #46 | 0.38 |
| Ítem #8 | 0.34 | Ítem #21 | 0.28 | Ítem #34 | 0.43 | Ítem #47 | 0.67 |
| Ítem #9 | 0.28 | Ítem #22 | 0.34 | Ítem #35 | 0.24 | Ítem #48 | 0.38 |
| Ítem #10 | 0.16 | Ítem #23 | 0.44 | Ítem #36 | 0.13 | Ítem #49 | 0.41 |
| Ítem #11 | 0.29 | Ítem #24 | 0.24 | Ítem #37 | 0.02 | Ítem #50 | 0.52 |
| Ítem #12 | 0.32 | Ítem #25 | -0.10 | Ítem #38 | 0.02 | Ítem #51 | 0.23 |
| Ítem #13 | 0.25 | Ítem #26 | 0.23 | Ítem #39 | 0.23 | Ítem #52 | 0.40 |

Fuente: Autores.

De acuerdo con los resultados en la medición del IF, se aislaron 6 ítems de la categoría Muy Difícil (ítems #4, #7, #25, #32, #36 y #37) y 3 ítems de la categoría Muy Fácil (ítems #1, #2 y #33), para su determinar si tenían posibilidad de corrección o si era necesario descartarlos. Además, según la medición del ID y teniendo en cuenta que los ítems con valores cercanos a 0 representan un posible efecto negativo en los examinados, se aislaron 8 ítems (#2, #5, #29, #31, #32, #36, #37 y #38) para su posterior revisión y posible corrección. El ítem #10 de valor negativo fue descartado completamente.

## Discusión

Primero se analizaron los 3 ítems de la categoría *Muy Fácil* (ítems #1, #2 y #33). Los dos primeros ítems corresponden a preguntas de comprensión oral de nivel A1, mientras que el ítem #33 corresponde a comprensión de lectura de nivel A1. Por lo tanto, en una población de la que se espera haya alcanzado el nivel A1 en inglés en su vida escolar reciente, es normal que los primeros ítems de nivel A1 aparezcan resueltos de manera correcta por la mayoría de examinados. El hecho de que los valores de IF de estos ítems no alcance el rango máximo de 1.00 punto quiere decir que, aun siendo muy fáciles, hubo una porción de la población examinada que falló en dar la respuesta acertada. De acuerdo con Toksöz y Ertunç (2017), estos ítems funcionan como "ejercicios de calentamiento" para la prueba

y generan un sentimiento inicial de logro y positivismo; por esta razón, además del hecho de que son una cantidad menor con respecto a todos los ítems de examen, se mantuvieron en el prototipo final de la prueba, y no se necesitó corrección para estos tres ítems.

Asimismo, se hizo revisión de los ítems de la categoría *Muy Difícil* (ítems #4, #7, #25, #32, #36 y #37). Se encontró que, si bien estos ítems suponen un reto alto para los examinados, hubo un porcentaje considerable de la población examinada que resolvió las preguntas correctamente, lo cual indica que no se trata de ítems imposibles de ser resueltos. De acuerdo con la literatura revisada, y teniendo en cuenta que se trata de una cantidad baja, se mantuvieron estos ítems en el prototipo final de la prueba, dado que este tipo de ítems jalona el desempeño de los examinados con mayor nivel y no genera efecto *washback* negativo en la población examinada (Brown, 2004; Toksöz & Ertunç, 2017). En general, se trata de una prueba exigente, como se aprecia en la Tabla 5:

Tabla 5.
*Síntesis de Interpretación del Índice de Dificultad (IF)*

| Rangos | Valoración | Cantidad de Ítems en UVEPLAT |
|---|---|---|
| 0.20 - menos | Muy difícil (considerar revisión o descartar) | 6 ítems (11%) |
| 0.21 - 0.40 | Difícil | 21 ítems (41%) |
| 0.41 - 0.60 | Moderado | 16 ítems (31%) |
| 0.61 - 0.80 | Fácil | 6 ítems (11%) |
| 0.81 - más | Muy fácil (considerar revisión o descartar) | 3 ítems (6%) |

Fuente: Autores.

La Tabla 5 permite ver que el 52% de los ítems están entre las categorías de *Difícil o Muy Difícil*, lo cual indica que UVEPLAT es una prueba con tendencia a la alta exigencia. En este punto es necesario comentar que en las etapas iniciales del diseño de UVEPLAT, uno de los principales retos fueron las creencias de los docentes diseñadores respecto a la evaluación; en general, todos creían que un buen examen era sinónimo de un examen difícil. Si bien el entrenamiento previo al diseño de la prueba se enfocó en desmantelar tales argumentos y crear conciencia de que una buena prueba es aquella que brinda al examinado suficientes oportunidades de demostrar su conocimiento, el diseño final de UVEPLAT y su exigencia podrían estar relacionados con las creencias previas de los diseñadores.

Con respecto a la revisión de los ítems con un potencial de discriminación muy pobre (ítems #2, #5, #29, #31, #32, #36, #37 y #38) se pudo establecer tres tipos de errores que pueden haber incidido en esos valores bajos:

- Se detectaron instrucciones ambiguas en el enunciado del ítem, dando lugar a más de una interpretación posible
- Se identificaron instrucciones demasiado largas, con explicaciones redundantes o con un lenguaje más complejo del que se pretende evaluar
- Se identificaron distractores ambiguos, de manera que más de una opción podría considerarse correcta

- Por último, se encontraron distractores mal diseñados, los cuales podían ser descartados fácilmente por los examinados sin tener conocimiento del tema evaluado.

De acuerdo con Palés-Argullós (2010), estos errores comunes en el diseño de ítems de respuesta múltiple deben evitarse al máximo. Al respecto, cabe mencionar que en las sesiones de estudio con el equipo de diseño se logró establecer una correlación entre la falta de entrenamiento en diseño de materiales y la falta de experiencia para diseñar instrucciones, distractores y tareas para evaluación. Si bien la evaluación es inherente al oficio diario del docente (así como la creación de materiales), los docentes que participaron como diseñadores de la prueba manifestaron carencias de entrenamiento formal en ambos ámbitos; estas carencias han sido documentadas en Colombia por autores como Núñez y Téllez (2009), Giraldo (2018a, 2018b, 2019) y López y Bernal (2009); de allí la importancia de que los docentes de inglés como lengua extranjera tengan mayor participación en el diseño de instrumentos de evaluación y de materiales para la enseñanza, toda vez que dichos diseños benefician a la institución y a los estudiantes, y empodera a los docentes (Núñez & Téllez, 2009). La Tabla 6 sintetiza la medición del índice de discriminación ID en UVEPLAT:

Tabla 6.
*Síntesis de Interpretación del Índice de Discriminación (ID)*

| Rangos | Valoración | Cantidad de Ítems en UVEPLAT |
|---|---|---|
| 0.40 - más | Ítem Excelente | 12 ítems (23%) |
| 0.30 - 0.39 | Ítem Bueno | 12 ítems (23%) |
| 0.20 - 0.29 | Ítem Aceptable | 18 ítems (35%) |
| 0.09 - 0.19 | Ítem Deficiente | 10 ítems (19%) |

Fuente: Autores.

La cantidad de ítems encontrados defectuosos en su capacidad discriminativa fue relativamente baja (19%) si se tiene en cuenta que 35% de los ítems son aceptables y 46% restantes están en las categorías de ***Bueno*** y ***Excelente***. Este resultado positivo es un producto directo de los esfuerzos en formación, producción e investigación de un grupo de docentes en torno al tema de evaluación y diseño de pruebas, así como de su participación en las discusiones, reflexiones y correcciones constantes al primer prototipo de UVEPLAT.

La medición del ID permitió la selección y corrección de 10 ítems. Se redactaron las instrucciones, de manera que fueran cortas, claras y precisas. De la misma manera, se descartaron distractores en todos estos ítems y se diseñaron distractores nuevos. Todas las correcciones fueron incorporadas al banco de ítems creado en Moodle. Vale la pena resaltar lo que afirman Taylor y Geranpayeh (2011) cuando dicen que "las pruebas son hasta cierto punto provisionales, trabajos en proceso, incluso experimentales, que con suerte sirven a un propósito práctico en el mundo real del aquí y el ahora" (p. 94), por lo tanto se espera continuar haciendo estudios y ajustes sobre la misma prueba.

### *Limitaciones y posibilidades de estudios futuros*

La principal limitación de este estudio tiene que ver con el hecho de que no se midió el índice de eficiencia en los distractores, entendido como la distribución de respuestas en cada uno de los distractores para determinar la frecuencia de escogencia de cada uno de ellos. Dado que se encontraron varios problemas relacionados con el diseño de distractores, valdría la pena un estudio más profundo sobre su comportamiento y su distribución en todos los ítems que componen la prueba UVEPLAT.

Futuros estudios deben proponerse alrededor de la validez y confiabilidad de la prueba. Al tratarse de un examen nuevo que empieza a implementarse y del que se derivan decisiones administrativas y académicas importantes, será necesario establecer la validez del contenido y de la prueba misma (content vailidity – face validity) y la confiabilidad de los resultados que arroja la prueba; el modelo que reportan Wall et al. (1994) resultaría muy pertinente para dichos estudios.

## Conclusiones

El análisis de ítems permitió establecer que UVEPLAT es una prueba funcional que cumple con su propósito de establecer el nivel de conocimiento del inglés como lengua extranjera, y que permitir la clasificación de estudiantes de primer semestre en los diferentes niveles que ofrece la Universidad. Luego de un año de trabajo con el equipo de diseño, el análisis de ítems permitió una mirada más clara de la calidad de cada pregunta diseñada, y proponer correcciones o cambios según los valores de sus índices de discriminación y de dificultad.

El ejercicio de diseñar un examen y dar inicio a pruebas estadísticas del mismo es una actividad enriquecedora tanto para el equipo de diseño como para la institución en general; el diseño de UVEPLAT implicó el estudio y discusión de referentes teóricos clásicos y contemporáneos sobre la evaluación de lenguas extranjeras, así como el estudio detallado de varios exámenes en plataformas comerciales y el análisis profundo del Marco Común Europeo de Referencia para las Lenguas. Todo esto reconectó a los docentes a una bibliografía esencial que podría convertirse en contenidos para el desarrollo profesional. Este proceso de estudio redundó en el empoderamiento de los docentes y en una contribución directa a su formación continua en términos de una mayor literacidad en la evaluación del inglés como lengua extranjera. Esto último se traduce en una mayor experiencia y diversificación de técnicas de evaluación, así como en el desarrollo de la reflexión constante y el pensamiento crítico hacia los procesos evaluativos. Por otra parte, con respecto a la institución, esta experiencia permitió un análisis detallado de las prácticas evaluativas en torno a las lenguas (materna y extranjera) que reveló varias deficiencias: el uso no sistemático de pruebas comerciales para propósitos múltiples dentro de la institución, la falta de formación en cuestiones de evaluación a nivel de pregrado, así como la falta de recursos financieros y de tiempo que se otorgan para la investigación en diseño de pruebas.

Se recomienda impulsar el desarrollo de más estudios empíricos que redunden en experiencia práctica e investigativa sobre el diseño de exámenes y la evaluación de los mismos

procesos evaluativos, el estudio estadístico de pruebas y el mejoramiento constante de tales procesos en el contexto local. En ese sentido, son bienvenidos todos los estudios empíricos locales, a nivel institucional, departamental y nacional, sobre temáticas de evaluación, de manera que la comunidad académica en nuestro país conformada por los docentes de lenguas extranjeras continúe su formación y crecimiento en LAL, contribuyendo así a un campo c¿en el que actualmente hay pocas investigaciones. Igualmente, se recomienda hacer la divulgación de estos estudios en español, pues no hay suficiente bibliografía sobre el tema en nuestra lengua, ni siquiera en los medios locales de circulación del conocimiento.

## Referencias


Bachman, L. F. (2004). *Statistical Analyses for language Assessment*. Cambridge: Cambridge University Press.

Brown, J. D. (2011). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment New Edition*. New York: McGraw-Hill.

Brown, D. (2004). *Language Assessment - Principles and Classroom Practice*. White Plains: Pearson education.

Carr, N. (2011). *Designing and Analyzing Tests*. Oxford: Oxford University Press.

Ebel, R. L. y Frisbie, D. A. (1986). *Essentials of Education Measurement*. Englewood Cliffs: Prentice Hall.

Fernández, M. (2007). Propuesta Metodológica para la Creación de un Nuevo Examen de inglés en las Pruebas de Acceso a la Universidad. [Tesis Doctoral]. Granada: Editorial de la Universidad de Granada.

Giraldo, F. (2018a). Language assessment literacy: Implications for language teachers.*Profile: Issues in Teachers' Professional Development*, 20(1), 179–195. http://doi.org/10.15446/profile.v20n1.62089

Giraldo, F. (2018b). A Diagnostic Study on Teachers' Beliefs and Practices in Foreign-Language Assessment. *Íkala, Revista de Lenguaje y Cultura, 23*(1), 25–44. http://doi.org/10.17533/udea.ikala.v23n01a04

Giraldo, F. (2019). Designing Language Assessments in Context: Theoretical, Technical, and Institutional Considerations. *HOW Journal, 26*(2), 123–143. https://doi.org/10.19183/how.26.2.512

ILTA. (2001). Código Ético. [*Online*]. Disponible en https://www.isa-sociology.org/es/sobre-isa/codigo-etico-440

Inbar-Lourie, O. (2013). Guest Editorial to the special issue on language assessment literacy. *Language Testing, 30*(3), 301–307. https://doi.org/10.1177/0265532213480126

Janssen, G. & Meier, V. (2013). Establishing placement test fit and performance: Serving local needs. *Colombian Applied Linguistics Journal, 15*(1), 100–113. https://doi.org/10.14483/udistrital.jour.calj.2013.1.a07

López, A. & Bernal, R. (2009). Language testing in Colombia: A call for more teacher-education and teacher training in language assessment. *Profile: Issues in Teachers' Professional Development, 11*(2), 55–70. Available from http://www.bdigital.unal.edu.co/16543/1/11442-28040-1-PB.pdf

Malau-Aduli, B. S. & Zimitat, C. (2012). The Analysis of Multiple-Choice Items of the Test of an Introductory Course in Chemistry in a Nigerian University. *International Journal of Learning*, 18(4), 237–246. https://doi.org/10.18848/1447-9494/CGP/v18i04/47579

Núñez, A. & Téllez, M.F. (2009). ELT Materials: The Key to Fostering Effective Teaching and Learning Settings. *Profile: Issues in Teachers' Professional Development, 11*(2), 171–186. Available from https://revistas.unal.edu.co/index.php/profile/article/view/11449

Palés-Argullós, J. (2010). ¿Cómo elaborar correctamente preguntas de elección múltiple? *Educación Médica, 13*, 149–155. https://doi.org/10.4321/S1575-18132010000300005

Paltridge, B. (1992). EAP placement testing: an integrated approach. *English for Specific Purposes, 11*, 243–268. https://doi.org/10.1016/S0889-4906(05)80012-2

Sugianto, A. (2020). Item Analysis of English Summative Test: EFL Teacher-Made Test. *Indonesian EFL Research & Practice*, 1(1), 35–54. Disponible en http://journal.ahsanta.ac.id/index.php/EFL/article/view/13

Taylor, L. & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10*, 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Toksöz, S. & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literacy Studies, 8*(6), 141–146. http://dx.doi.org/10.7575/aiac.alls.v.8n.6p.141

Tomlinson, B. (2005). Testing to learn: a personal view of language testing. *ELT Journal, 59*(1), 39–46. https://doi.org/10.1093/elt/cci005

Wall, D., Claphman, C. & Alderson, C. (1994). Evaluating a placement test. *Language Testing, 11*, 321–344. https://doi.org/10.1177/026553229401100305

**Alexánder Ramírez Espinosa** es profesor de Inglés y Lingüística en la Escuela de Ciencias del Lenguaje, de la Universidad del Valle. Es licenciado en Lenguas Extranjeras y Magíster en Lingüística, ambos títulos obtenidos en la Universidad del Valle. Recientemente ha iniciado sus estudios de Doctorado en Educación con énfasis en ELT Education. Alexánder imparte sus cursos de inglés y lingüística en la Licenciatura e Lenguas Extranjeras, en la Maestría en Estudios Interlingüísticos e Interculturales y en la Tecnología en Interpretación para Sordos y Sordociegos. Sus intereses investigativos incluyen la evaluación en lenguas extranjeras, la comunicación intercultural, el desarrollo de la autonomía del aprendiz y la Lingüística Queer. https://orcid.org/0000-0002-7122-9537