QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Validación de Pruebas de Inglés EFL de Bajo Riesgo

Pablo Garcés[1]

Recibido 17 de diciembre de 2019

Aceptado 27 de diciembre de 2019

Resumen: Ya sea que un candidato esté presentando una prueba de "alto" riesgo como el TOEFL o una de "bajo" riesgo, como el examen de final de un nivel, las repercusiones de una calificación errada siempre tienen un impacto importante. Es por eso que las pruebas de alto riesgo son constantemente sometidas a procesos de validación, mientras que sus contrapartes de bajo riesgo rara vez lo son. Este trabajo de investigación ha dirigido su atención a un curso de inglés EFL de una institución panameña con el objetivo de determinar la validez de las pruebas utilizadas en el programa, consideradas de "bajo" riesgo. El tipo de investigación realizado es aplicada, descriptiva, mixta, no experimental y transversal. Se evaluó: la validez basada en el contexto, la validez basada en la teoría, la confiabilidad de los resultados y las características de los candidatos. Se estudiaron las pruebas institucionales del curso de inglés EFL de una institución panameña. En la misma, la mayoría de los profesores utilizan la plataforma LMS Schoology para aplicar sus exámenes de fin de curso. Se tomó una muestra en junio de 2018 que contenía el registro de las pruebas finales de los niveles 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13 y 15; ya que estos son los niveles cuyos profesores utilizaron la plataforma en ese período. La muestra contenía un total de 340 aplicaciones. Se encontraron importantes deficiencias en la evaluación de

[1]Profesor de Introducción a la Comunicación en Quality Leadership University. Posee una licenciatura en Comunicación Social de la Universidad Rafael Belloso Chacín y una maestría en Ciencias de la Comunicación de la Universidad Rafael Belloso Chacín, Venezuela. **Correo electrónico** pablo.garces@qlu.pa

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

validez basada en el contexto y en la confiabilidad de los resultados; mientras que en la evaluación de validez basada en teoría y la de las características de los candidatos se obtuvieron resultados muy positivos. Se aportaron una serie de recomendaciones entre las cuales se encuentran: la creación de pruebas orales institucionales, la creación de una sección de evaluación auditiva y la profundización de los estudios de confiabilidad, entre otras.

Palabras Clave: Validez, Validación, Instrumentos de Evaluación, Pruebas de Inglés, EFL, Pruebas de bajo riesgo, Pruebas de alto riesgo.

Abstract: Whether a candidate is taking a "high stakes" test such as the TOEFL or a "low stakes" test such as an end-of-level test, the repercussions of a wrong score always have a significant impact. That's why high-stake tests are constantly subjected to validation processes, while their low-stakes counterparts rarely are. This research has focused its attention on an EFL program of a Panamanian institution with the objective of determining the validity of the tests used, which are considered to be "low" stakes. The type of research carried out is applied, descriptive, mixed, non-experimental and transversal. The following types of validity were evaluated: validity based on the context, validity based on the theory, reliability of test results and characteristics of the candidates. The institutional tests of the EFL courses of a Panamanian institution were studied. In this institution most of the teachers use the Schoology LMS platform to administer their end-of-course exams. A sample was taken in June 2018 that contained the record of the final tests of levels 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13 and 15; since only the teachers of these levels used the platform in that period. The sample contained a total of 340 administrations. Important deficiencies were found in regards to context-based validity and score reliability, while the results of the evaluation of theory-based validity and the characteristics of the candidates were rather positive. A series of recommendations were provided, among which are the creation of

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

institutional oral tests, the inclusion of a section to evaluate listening comprehension, and the application of greater reliability studies, among others.

Key words: Validity, Validation, Evaluation instruments, English tests, EFL, Low stakes tests, High stakes tests.

INTRODUCCIÓN

Las instituciones que enseñan inglés como segunda lengua dan gran importancia a los métodos de evaluación que son aplicados a los estudiantes, ya sean como método de diagnóstico de nivel de inglés, de evaluación continua o para aprobar un nivel o grado. Sin embargo, los métodos para crear dichos exámenes varían enormemente entre las diversas instituciones.

Es por eso que existen procesos de validación que pueden ser aplicados a las pruebas y a sus resultados para determinar qué tan confiables son. Es importante saber que los exámenes de inglés pueden separarse en dos categorías: “alto riesgo” y “bajo riesgo”. Las pruebas de alto riesgo son aquellas que son reconocidas internacionalmente como TOEFL o IELTS. Es común que éstas sean solicitadas por universidades y empresas internacionales como comprobante del dominio del idioma de un candidato. Éstas son frecuentemente sometidas a procesos de validación, lo cual hace que sus resultados sean muy confiables. Por otra parte, las pruebas de bajo riesgo son aquellas diseñadas de forma independiente por escuelas, profesores e institutos y son utilizadas frecuentemente para evaluar el desempeño de los estudiantes al finalizar un módulo, nivel o grado. Estas pruebas poseen un impacto sólo dentro del contexto en el que se encuentran, es decir, impactan en la nota del estudiante o determinan si pasa el nivel o si tiene que repetirlo. Coincidentemente, los autores Chapelle, Jamieson y Hegelheimer (2003) toman en consideración que, debido al efecto mínimo que tienen en lo que respecta a la toma de decisiones, las pruebas de bajo riesgo tienden a tener procesos menos rigorosos de validación, en caso de tener alguno.

Sin embargo, Kokhan (2013) argumenta que las pruebas de bajo riesgo resultan prácticas para las instituciones pero podrían tener un rango de error de hasta 40% debido a su nivel bajo de validez.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

En el mismo orden de ideas, Chapelle, et al. (2003) sugieren la posibilidad de someter a procesos de validación, inclusive a las pruebas de bajo impacto, ya que dicha categorización no toma en cuenta otras formas de impacto que dicho resultado puede tener en los estudiantes y se debe considerar que los resultados errados pueden crear una falsa sensación de progreso o desmotivación relacionado con niveles equivocados de desempeño.

Es por ello que se considera de suma importancia validar de forma sistemática las pruebas que se utilizan en la enseñanza del inglés como segunda lengua o como lengua extranjera y determinar qué tan precisas son en la medición de competencias comunicativas de los estudiantes de inglés, así como también, empezar a crear un antecedente de validación de pruebas de bajo riesgo. Se ha tomado un curso de inglés EFL de una institución panameña que implementó la creación de sus pruebas de bajo riesgo para la aprobación de niveles y se procedió a aplicar un proceso de validación y determinar en qué medida dichas pruebas son válidas.

Sería ideal poder realizar un proceso de validación a un grupo amplio de instituciones y a una cantidad elevada de pruebas que permitan detectar áreas de mejora a escala nacional o mundial. Sin embargo, realizar dicho procedimiento requiere acceso a resultados de estudiantes de dichas instituciones que generalmente compiten entre ellas, por lo que se vuelve logísticamente difícil. Sin embargo, al realizar un estudio en una institución, se puede generar un antecedente importante que permita a futuros investigadores tener un marco de referencia para validar otras pruebas que sean consideradas de bajo riesgo de otras instituciones.

OBJETIVOS DE LA INVESTIGACIÓN

Objetivo General: Determinar la validez de las pruebas utilizadas en el curso de inglés EFL de una institución panameña con el fin de analizar la confiabilidad de sus resultados.

Objetivos Específicos:

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

- Analizar el contenido de las pruebas institucionales para aportar una medida de validez.
-

- Analizar los resultados de las pruebas aplicadas para medir el grado de confiabilidad.
- Analizar las condiciones en las cuales se aplican los métodos de evaluación orales y escritos aplicados en la enseñanza de inglés como segunda lengua.

DELIMITACIÓN

El estudio se desarrolló en una institución panameña que ofrece cursos de inglés EFL. Se tomó en cuenta las aplicaciones de las pruebas hechas a través de la plataforma LMS Schoology, ya que la mayoría de los profesores la utilizan para aplicar sus exámenes de fin de curso. Por ello, la muestra consiste en los resultados de los niveles 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13 y 15. Cabe destacar que los profesores de los niveles 8,10 y 14 no utilizaron la plataforma, por lo cual no fueron tomados en cuenta como parte del objeto de estudio. La muestra se tomó durante la semana de pruebas finales en junio de 2018 que resultaron ser un total de 340 aplicaciones.

MARCO TEÓRICO

Weir (2005) indica que "La validación de una prueba es el proceso de generar evidencia para respaldar la buena base de las inferencias sobre el rasgo de los puntajes de las pruebas". De esta manera, el autor hace referencia a la evidencia necesaria para realizar un argumento de validez, más allá de la intuición o la percepción subjetiva que se pueda tener sobre la validez de una prueba.

Weir (2005) además compara el procedimiento con un abogado defensor que actúa en el tribunal. De esta manera, este abogado debe presentar evidencia que concierne a la validación de constructo, definida por Brown (1996) como "el grado en que una prueba

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

mide lo que dice o pretende medir". Esto, según Weir (2005), puede medirse tomando en cuenta principalmente, tres aspectos de la evaluación: la validez basada en el contexto, la validez basada en la teoría y la confiabilidad de los resultados, tomando en cuenta, a su vez, otros elementos como lo son las características de los candidatos, entre otros.

La validez basada en el contexto.

Para Pennington (2003), la validez basada en contexto se refiere a la medida en que una medición representa todas las facetas de una construcción dada. Weir (2005) expande su definición de la siguiente manera: "La validez del contexto se refiere al grado en que la selección de tareas en una prueba es representativa del universo más amplio objetivos del cual se supone que la prueba es una muestra. Esta cobertura se relaciona con las demandas lingüísticas e interlocutoras hechas por la(s) tarea(s), así como las condiciones bajo las cuales se realiza la tarea, derivadas tanto de la tarea misma como de su entorno administrativo".

Cada tarea dentro de una prueba exige que el candidato demuestre una habilidad. La validez del contexto se encarga de demostrar que estas tareas son una representación de habilidades reales que se requiere que sean dominadas al momento de poner el inglés en práctica. O'Sullivan et al. (2002) realizaron un estudio en el que trataron de demostrar como el Cambridge ESOL realizaba esfuerzos por llevar a cabo una validación de contexto a través de una lista de verificación durante las pruebas orales. Así mismo, Weir (2005) asegura que los siguientes elementos deben ser analizados: Planteamiento de objetivos, propósito, formato de respuesta, criterios conocidos, distribución de puntuación, orden de ítems, restricciones de tiempo, condiciones físicas, uniformidad de administración, seguridad, modo de discurso, canal, extensión, naturaleza de la información, conocimiento del contenido, léxico, gramática y estructura.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

La Validez basada en la Teoría.

Para estudiar el argumento de validez basado en teoría se debe crear un marco de referencia para determinar cuáles son las habilidades comunicativas de lenguaje que se espera que los estudiantes desarrollen a lo largo del curso y que serán objeto de evaluación.

Bachman (1990) define al lenguaje comunicativo en los siguientes términos:

La capacidad del lenguaje comunicativo consiste en la competencia lingüística, la competencia estratégica y los mecanismos psicofisiológicos. La competencia lingüística incluye la competencia organizacional, (gramatical y textual) y la competencia pragmática, (locutorio y sociolingüístico). La competencia estratégica es la capacidad que relaciona la competencia del lenguaje con las estructuras de conocimiento del usuario del lenguaje y las características del contexto en el que tiene lugar la comunicación. La competencia estratégica realiza funciones de evaluación, planificación y ejecución para determinar los medios más efectivos para alcanzar un objetivo comunicativo. Los mecanismos psicofisiológicos implicados en el uso del lenguaje caracterizan el canal (auditivo, visual) y el modo (receptivo, productivo) en el que se implementa la competencia. (1990: 107)

Partiendo de lo argumentado por el autor, cada prueba debe contener, en cierta medida, elementos que permitan evaluar si los candidatos poseen el dominio adecuado de cada una de las competencias anteriormente mencionadas. Según Weir (2005), dichos elementos son los siguientes: planteamiento de objetivos, medición de dominio de lectura, contenido y temática, generación de ideas, organización de ideas, traducción de ideas, conocimiento gramatical de lenguaje, conocimiento textual de lenguaje, conocimiento funcional de lenguaje, conocimiento sociolingüístico de lenguaje y conocimiento de contenido.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

La confiabilidad de los resultados

Estudiando la "confiabilidad", Weir (2005) prefería referirse a dicho concepto bajo el término "validez de calificación", partiendo de las consideraciones que Alderson (1991) tenía con respecto a la complejidad que tenía la comprensión de la "confiabilidad" y todos sus aspectos. De acuardo a Weir (2005), "la validez de calificación se refiere a la medida en que los resultados de las pruebas son estables a lo largo del tiempo, consistentes en términos del muestreo de contenido y libres de sesgo". A su vez, Anastasi (1988) resalta que la variación de los resultados que un grupo de candidatos puedan tener a lo largo de múltiples examinaciones con diversas condiciones, permite separar los elementos a los que se le pueden atribuir posibilidad de error. En Standards for Educational and Psychological Testing, un conjunto de reglas elaborados en colaboración entre el American Educational Research Association, American Psychological Association, y el National Council on Measurement in Education (1974, 1985, 1999), las siguientes categorías de confiabilidad son ampliamente reconocidas: confiabilidad examinación-reexaminación; confiabilidad de formularios paralelos; consistencia interna; confiabilidad de marcador.

En esta investigación, la categoría de confiabilidad utilizada para la validación de la prueba fue la consistencia interna, por su habilidad de medir si varios ítems que proponen medir la misma construcción general producen puntajes similares.

Las características de los candidatos

O'Sullivan (2000) en su libro *Towards a Model of Performance in Oral Language Testing,* discernió sobre las características de los candidatos que podrían tener un efecto potencial en su desempeño en una prueba. Concluyó que estas eran las categorías y sus factores:

Físicas/Fisiológicas: enfermedades a corto plazo; discapacidades a largo plazo; edad y sexo

Sicológicas: personalidad, memoria, estilo cognitivo, esquema afectivo, concentración, motivación y estado emocional

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Experienciales: educación, preparación para examinación, experiencia en examinación y experiencia en comunicación.

Weir (2005) es un importante referente para este trabajo, ya que diseñó un marco de referencia desde indicando el procedimiento procedimientos más adecuado para la generación de un argumento de validez y dicho procedimiento fue el que se utilizó primordialmente para llevar a cabo la presente investigación.

METODOLOGÍA

El tipo de investigación realizado es aplicada, ya que se buscó utilizar los conocimientos en la práctica de forma que resulte a un aporte a la sociedad.

Adicionalmente, es descriptiva, ya que el autor buscó describir las propiedades y características de los fenómenos que están siendo sometidos a análisis y existió una necesidad de realizar un proceso de recolección de datos para alcanzar el objetivo, recaudando toda la información posible con respecto a la variable "métodos de evaluación", sus dimensiones e indicadores. Asimismo se presentó una relación de asociación entre las variables.

Por otra parte, es cuantitativa, ya que se establecieron mediciones reales para poder obtener datos estadísticos precisos de los "métodos de evaluación" que se pudieron verificar, contrastar y analizar con el propósito de encontrar patrones y datos objetivos relevantes. A su vez es cualitativa, ya que el análisis de resultados conllevará a un argumento de validez refiriéndose a los significados, características, definiciones, conceptos, símbolos y descripción de los elementos estudiados.

Fuentes de Información

Para la realización de este trabajo de investigación se revisó principalmente la base de datos EBSCO y Science Direct de la cual se obtuvieron una serie de trabajos científicos

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

que definen los elementos estructurales y cualidades de los instrumentos de evaluación en diferentes contextos.A través de Google Books se adquirió el libro Language Testing and Validation de C. Weir, el cual sirvió como guía principal para llevar a cabo el proceso de validación.

Se estudiaron las pruebas institucionales del curso de inglés EFL de una institución panameña. La misma, les permite a sus profesores utilizar la plataforma LMS Schoology

para aplicar sus exámenes de fin de curso. Se tomó una muestra en junio de 2018 que contenía el registro de las pruebas finales de los niveles 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13 y 15; ya que estos son los niveles cuyos profesores utilizaron la plataforma para aplicar la prueba final de este período. La muestra contenía un total de 340 aplicaciones.

Variables

1. Variable independiente: Pruebas EFL de bajo riesgo.

2. Variable dependiente: Validación de pruebas EFL de bajo riesgo.

Se adaptaron 2 instrumentos de recolección de datos tipo encuesta, partiendo de las categorías planteadas por Weir (2005) para ser aplicadas a estudiantes y profesores.

El instrumento #1, como se puede ver en la figura 1, consiste en una encuesta de 10 preguntas creadas en Google forms, dirigida a los estudiantes que se encuentran a punto de tomar una prueba. Se le pidió a cada candidato que completara la encuesta antes de dar inicio a su examen. Los ítems contenidos en el instrumento permitieron medir las características de los candidatos que deben ser tomadasen cuenta según O'Sullivan etal. (2002).

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 1.

Escala / Preguntas	Totalmente en desacuerdo	En desacuerdo	Ni acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo
1. No tengo ninguna molestia física (dolor de cabeza, resfriado, etc)					
2. No tengo ningún problema de visión que afecte mi desempeño					
3. No tengo alguna condición que pueda causarme confusión al responder (ej: dislexia, agrafia, disperflexia, disgrafia, disfasia, etc)					
4. Me siento confiado para tomar este examen					
5. Recuerdo bien cuáles son los temas que serán					

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

evaluados en este examen					
6. Opino que los exámenes siempre son importantes					
7. Me siento motivado a presentar este examen					
8. Tengo experiencia tomando cursos de inglés					
9. Estudié lo necesario para tomar este examen					
10. Tengo experiencia presentando exámenes de inglés de selección simple					

Fuente (autoría del investigador)

Por otra parte, el instrumento #2, como se puede ver en la figura 2, consiste en una encuesta de 34 preguntas, también creada en Google forms, la cual está dirigida a los profesores encargados de aplicar cada prueba y que debían diligenciar una vez que lo habían hecho. Fueron un total de 12 profesores. También se les pidió que accedieran a la misma a través de un link. Los ítems de la encuesta permitieron medir la presencia de los elementos de validez basado en contexto según Weir (2005) y la validez basado en teoría según de Bachman y Palmer (1996).

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 2.

Escala / Preguntas	Totalmente en desacuerdo	En desacuerdo	Ni acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo
1. El examen explica a los candidatos exactamente lo que deben hacer en cada pregunta					
2. El examen explica lo que el candidato puede o no hacer					
3. El examen explica de forma inequívoca el propósito del mismo					
4. El formato es el adecuado para que no afecte el desempeño					
5. El criterio de evaluación está explicado de forma clara					
6. El peso de cada componente de la prueba está alocado de forma adecuada					
7. El orden de los Ítems es el adecuado					

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

8. El tiempo para completar el examen es apropiado					
9. Las condiciones son adecuadas (temperatura, silencio, iluminación)					
10. Las condiciones son iguales en cada aplicación del examen					
11. Hay seguridad de que las pruebas no han sido copiadas o publicadas					
12. La prueba examina la habilidad de escritura de los candidatos					
13. La prueba examina la habilidad de lectura de los candidatos					
14. La prueba incluye una evaluación oral estándar					
15. La prueba examina la habilidad de escucha de los candidatos					
16. La prueba escrita es un canal apropiado para medir si los objetivos del nivel fueron logrados					
17. El largo de la prueba es apropiado para					

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

medir si los objetivos del nivel fueron logrados					
18. El tipo de información usado en la prueba (el tema y los ejemplos) es apropiado para medir si los objetivos del nivel fueron logrados					
19. Los temas utilizados en la prueba son adecuados a la edad y experiencia de cada estudiante					
20. El examen está libre de elementos que generen incomodidad o controversia (guerra, religión, política)					
21. El vocabulario del examen está de acuerdo al nivel de los candidatos					
22. La gramática del examen está de acuerdo al nivel de los candidatos					
23. El examen comprueba de forma satisfactoria la habilidad de comunicarse de los					

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

candidatos					
24.La prueba permite evaluar la habilidad que tiene el candidato de leer superficialmente un texto					
25.La prueba permite evaluar la habilidad que tiene el candidato de localizar información específica en un texto					
26.La prueba contiene temas relevantes					
27.La prueba fomenta la producción de ideas por parte del candidato					
28.La prueba invita al candidato a organizar y categorizar ideas					
29.La prueba invita al candidato a poner las ideas en un lenguaje apropiado, cohesivo y coherente					
30.La prueba permite evaluar la habilidad gramática que tiene el candidato					

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

31.La prueba permite evaluar la habilidad que tiene el candidato de entender la cohesión y coherencia del texto					
32.La prueba permite evaluar la habilidad que tiene el candidato de entender el lenguaje de un entorno socio-cultural determinado					
33.La prueba comprueba conocimiento previo que tiene el candidato del contenido					
34.La prueba aporta conocimiento nuevo al candidato					

Fuente (autoría del investigador)

Por último, se registraron los resultados de todos los estudiantes en un archivo simple de Excel de todas las pruebas aplicadas, como se puede ver en la figura 4, con la intención de aplicar la fórmula estadística Kuder-Richardson 20 (KR20) para determinar la consistencia interna de los resultados para cada prueba. Según Kuder y Richardson (1937), la fórmula mide la confiabilidad de consistencia interna para instancias que presentan opciones dicotómicas, en otras palabras, cumple la misma función del Alpha de Cronbach calculado para puntajes dicotómicos. Cada pregunta contestada de forma correcta se contabilizó mediante la medida 1 y aquellas contestadas de forma incorrecta se le adjudicaron un 0. A los resultados se le aplicó la fórmula reflejada en la figura 3.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Fórmula 1.

$$Kr20 = \frac{K}{K-1}\left\{1-\left\{\frac{\sum \sigma^2}{\sigma^2}\right\}\right\}$$

Los resultados se ponderaron de la siguiente manera:

Coeficiente KR20 mayor a 0,9 es Excelente

Coeficiente KR20 mayor a 0,8 y menor a 0,9 es Bueno

Coeficiente KR20 mayor a 0,7 y menor a 0,8 Aceptable

Coeficiente KR20 mayor a 0,6 y menor a 0,7 Cuestionable

Coeficiente KR20 mayor a 0,5 y menor a 0,6 Pobre

Coeficiente KR20 menor a 0,5 es Inaceptable

Cuadro 3.

Nivel: _____

Estudiantes/Preguntas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Estudiante 1																																																			
Estudiante 2																																																			
Estudiante 3																																																		
Estudiante 4																																																		
Estudiante 5																																																		
Estudiante 6																																																		
Estudiante 7																																																		
Estudiante 8																																																		
Estudiante 9																																																		
Estudiante 10																																																		
Estudiante 11																																																		
Estudiante 12																																																		
Estudiante 13																																																		
Estudiante 14																																																		
Estudiante 15																																																		
Estudiante 16																																																		
Estudiante 17																																																		
Estudiante 18																																																		
Estudiante 19																																																		
Estudiante 20																																																		

Fuente (autoría del investigador)

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

RESULTADOS Y DISCUSIÓN

Los resultados de la aplicación de los instrumentos #1 y #2 se tabularon (Figuras 5, 6, 7 y 8) para ser analizados en función de los indicadores. En dichos resultados se estableció una medida del 1 al 5, en la cual las medidas más cercanas al uno, resaltadas con una tonalidad más roja, demuestran condiciones de los candidatos menos adecuadas para la obtención de resultados precisos, mientras que las medidas más cercanas al 5, resaltadas con una tonalidad más verde, demuestran condiciones más adecuadas.

Resultados de Instrumento #1:

Como se puede ver en la figura 5, entre los resultados más relevantes podemos destacar la prueba de nivel 2 la cual refleja medidas bajas dentro de los ítems 1, 2 y 3, los cuales se encuentran relacionados con inconvenientes de carácter físico que estuvieron presentes durante las pruebas. Este dato podría correlacionarse con los resultados de dichas pruebas y en caso de haber problemas de rendimiento estos se le podrían atribuir a dichos inconvenientes o por lo menos argumentar que pudo haber tenido un impacto en su desempeño durante la prueba y consecuentemente en los resultados obtenidos en los mismos. Lo anterior influencia la cualidad de validez de los resultados.

Otro ejemplo observable es como los estudiantes de los cursos de niveles intermedios (niveles 6, 7, 8 y 9) muestran valores altos en los ítems 4, 5 y 6, los cuales se refieren a la predisposición que los candidatos pueden tener a la prueba, lo cual demuestra, en dichos niveles, buena disposición para tomar la prueba, lo cual, por el contrario, aumentaría la validez de los resultados.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 5.

Nivel/Pregunta	1	2	3	4	5	6	7	8	9	10
Nivel 1	3.5	4.5	4.5	3.5	3.9	4.5	3.8	2	2.9	2.9
Nivel 2	2.4	2.1	2.4	3.9	4.1	4	3.9	2.9	3.5	3.5
Nivel 3	4.4	4.4	4.8	4.1	4.2	4.3	4.4	3.2	3.7	3.9
Nivel 4	4.5	4.5	4.6	3.8	4.1	4.6	3.9	3.4	3.5	4.2
Nivel 5	4.6	4.6	5	4.9	4.6	4.3	4.4	2.7	4.3	3.7
Nivel 6	4.8	4.7	4.8	4.7	4.6	4.3	4.1	3.7	3.8	4.3
Nivel 7	3.3	3.3	3.3	4.1	4.4	4.5	4.4	3.6	3.5	4.5
Nivel 9	4.8	4.6	4.9	4.6	4.7	4.9	4.7	4	4.7	4.3
Nivel 11	3.8	4.4	4.3	4.3	4.4	4.3	3.9	4	3.6	4
Nivel 12	4.8	4.8	4.6	4.2	4.3	4.7	4.1	4.2	3.3	4.4
Nivel 13	3	3.4	3.3	3.4	4	4.1	4.3	4.3	3.4	4.7
Nivel 15	3.6	3.2	4.8	3.2	3	4.4	2.8	3.2	2.6	3.2
No indicó	4.1	4.2	4.5	4.2	4.4	4.5	4.1	2.5	4	2.9

Fuente: autoría del investigador

Resultados de Instrumento #2:

Los resultados muestran deficiencias importantes en elementos clave del argumento de validez que pueden tener las pruebas. En la figura 6 se puede observar que el resultado más alarmante es la deficiencia de los ítems 14 y 15 en casi todas las pruebas. Estos ítems están relacionados con la evaluación oral (ítem 14) y de escucha de los candidatos (ítem 15); esto no significa que el programa no posea una evaluación para dichos elementos, ya que se pudo comprobar que sí existe una evaluación oral, pero sí quiere decir que dicha evaluación no se encuentra estandarizada y varía de curso en curso, afectando directamente la validez de sus resultados.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 6

	Item 14	Item 15
Indique el nivel que corresponde a la prueba	La prueba incluye una evaluación oral estándar	La prueba examina la habilidad de escucha de los candidatos
Level 1	1.0	1.0
Level 2	1.3	1.7
Level 3	1.7	1.0
Level 4	1.3	1.3
Level 5	1.3	1.3
Level 6	1.7	1.3
Level 7	1.7	1.7
Level 8	1.3	1.0
Level 9	1.7	1.3
Level 10	1.3	2.0
Level 11	1.0	1.0
Level 12	2.3	1.7
Level 13	2.3	1.3
Level 14	1.0	1.7
Level 15	1.7	1.3

Fuente: Autoría del investigador

Otros elementos que muestran importantes fallas son los ítems 1, 2 y 3, observables en la figura 7, los cuales se enfocan en la orientación que la prueba debe dar a los candidatos para que estos estén al tanto de cuáles son los objetivos y los reglamentos que deben seguir a lo largo de la prueba. Según los resultados actuales, se puede inferir que las pruebas tienen muy poca descripción que guíe a los candidatos. Esto podría indicar que los resultados de las pruebas podrían haber sido impactados por la incomprensión de los candidatos de los objetivos o los reglamentos de la evaluación en sí.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 7.

	Item 1	Item 2	Item 3
Indique el nivel que corresponde a la prueba	El examen explica a los candidatos exactamente lo que deben hacer en cada pregunta	El examen explica lo que el candidato puede o no hacer	El examen explica de forma inequivoca el proposito del mismo
Level 1	1.3	2.0	2.7
Level 2	2.3	3.0	4.0
Level 3	2.3	2.3	3.0
Level 4	3.0	3.7	4.0
Level 5	2.7	2.7	2.7
Level 6	3.7	2.0	2.0
Level 7	2.7	3.3	2.3
Level 8	3.7	3.3	4.0
Level 9	3.3	3.0	2.3
Level 10	3.0	2.3	3.0
Level 11	3.0	3.3	3.0
Level 12	3.7	4.0	3.0
Level 13	2.7	2.7	3.3
Level 14	3.0	2.7	3.3
Level 15	1.3	1.3	1.3

Fuente: Autoría del investigador

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Sin embargo, existen otros elementos que aportan niveles adecuados de validación a la prueba: los ítems 6, 7, 8, 9 y 10, (ver figura 8) que se encargan de medir las condiciones de la prueba (temperatura, ruido, tiempo, orden de preguntas, consistencia en aplicación). Los mismos obtuvieron resultados de la medición entre aceptables y adecuados. Esto indica que las condiciones generales en las cuales se toma el examen son adecuadas ya que no interfieren, o interfieren poco, con el desempeño de los candidatos.

Cuadro 8.

	Item 6	Item 7	Item 8	Item 9	Item 10
Indique el nivel que corresponde a la prueba	El peso de cada componente de la prueba está alocado de forma adecuada	El orden de los Items es el adecuado	El tiempo para completar el examen es apropiado	Las condiciones son adecuadas (temperatura, silencio, iluminación)	Las condiciones son iguales en cada aplicación del examen
Level 1	3.0	2.7	3.0	4.7	4.7
Level 2	3.3	4.0	3.3	3.3	4.3
Level 3	3.3	3.7	3.3	3.7	5.0
Level 4	4.3	3.3	4.0	4.0	4.3
Level 5	3.7	3.7	3.7	3.3	3.7
Level 6	3.7	3.7	3.7	4.0	4.0
Level 7	3.7	4.3	3.3	3.0	3.7
Level 8	2.7	4.7	2.7	2.7	4.0
Level 9	2.7	3.0	4.0	4.0	3.7
Level 10	3.7	3.3	3.7	2.7	3.3
Level 11	4.3	3.7	4.3	4.7	4.3
Level 12	4.7	3.3	4.0	4.3	3.7
Level 13	4.3	4.0	4.7	4.3	4.0
Level 14	3.7	4.0	4.0	3.7	4.0
Level 15	4.0	3.0	4.3	3.3	2.3

Fuente: Autoría del investigador

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Como se puede ver en las figuras 9 y 10, los resultados de los ítems 13, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33 y 34 fueron muy positivos con excepción del ítem 29. Éstos se encargan de estudiar la capacidad de la prueba de medir las habilidades gramática, lectura, habilidades comunicativas, vocabulario, alcance de objetivos y manejo de contenido. Este resultado indica que en gran parte, las pruebas son exitosas evaluando en qué medida los candidatos aprendieron el contenido que la institución esperaba que aprendiera y ayuda a construir un argumento a favor de la validez de las pruebas.

Cuadro 9.

	Item 13	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24
Indique el nivel que corresponde a la prueba	La prueba examina la habilidad de lectura de los candidatos	La prueba escrita es un canal apropiado para medir si los objetivos del nivel fueron logrados	El largo de la prueba es apropiado para medir si los objetivos del nivel fueron logrados	usado en la prueba (el tema y los ejemplos) es apropiado para medir si los objetivos del nivel fueron logrados	Los temas utilizados en la prueba son adecuados a la edad y experiencia de cada estudiante	El examen está libre de elementos que generen incomodidad o controversia (guerra, religión, política)	El vocabulario del examen está de acuerdo al nivel de los candidatos	El la gramática del examen está de acuerdo al nivel de los candidatos	El examen comprueba de forma satisfactoria la habilidad de comunicarse de los candidatos	La prueba permite evaluar la habilidad que tiene el candidato de leer superficialmente un texto
Level 1	4.3	3.7	3.3	3.7	3.0	5.0	2.7	3.7	3.0	2.7
Level 2	5.0	3.7	3.7	3.3	2.7	4.3	5.0	4.7	3.0	4.3
Level 3	4.3	3.7	3.7	3.3	2.3	4.3	4.3	3.0	3.0	4.3
Level 4	4.0	4.0	4.0	4.0	4.3	4.7	4.7	4.7	4.7	3.7
Level 5	4.7	2.7	3.7	3.7	2.7	4.3	4.7	5.0	3.3	3.7
Level 6	4.3	4.3	3.7	3.3	3.3	3.7	4.7	4.3	3.3	4.0
Level 7	4.0	3.7	3.7	3.0	2.3	4.0	4.7	4.0	3.0	4.0
Level 8	4.7	3.7	3.7	3.3	2.3	4.3	4.3	4.3	3.0	2.7
Level 9	3.7	3.3	3.7	2.7	2.3	3.7	4.0	4.0	2.7	2.7
Level 10	4.0	2.0	3.7	3.7	3.3	4.3	4.7	4.3	3.3	4.0
Level 11	4.7	3.3	3.7	3.7	4.0	4.7	4.7	4.7	4.0	4.3
Level 12	4.0	3.7	3.7	3.3	4.3	4.3	4.0	3.3	3.3	4.3
Level 13	4.7	3.0	4.3	4.3	4.0	4.3	4.7	5.0	4.0	4.3
Level 14	5.0	3.3	3.7	3.7	3.0	4.7	5.0	5.0	3.0	4.3
Level 15	3.7	2.0	2.3	3.0	3.3	4.0	3.7	3.7	2.7	1.7

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 10.

	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30	Item 31	Item 32	Item 33	Item 34
Indique el nivel que corresponde a la prueba	La prueba permite evaluar la habilidad que tiene el candidato de localizar información específica en un texto	La prueba contiene temas relevantes	La prueba fomenta la producción de ideas por parte del candidato	La prueba invita al candidato a organizar y categorizar ideas	La prueba invita al candidato a poner las ideas en un lenguaje apropiado, cohesivo y coherente	La prueba permite evaluar la habilidad gramática que tiene el candidato	La prueba permite evaluar la habilidad que tiene el candidato de entender la cohesión y coherencia del texto	evaluar la habilidad que tiene el candidato de entender el lenguaje de un entorno socio-cultural determinado	La prueba comprueba conocimiento previo que tiene el candidato del contenido	La prueba aporta conocimiento nuevo al candidato
Level 1	2.7	3.0	2.0	2.0	2.7	2.0	3.3	3.3	3.7	4.3
Level 2	3.7	4.0	3.7	3.7	2.7	4.7	3.0	2.7	4.0	3.0
Level 3	3.7	4.0	3.7	3.7	2.7	4.7	3.7	2.7	4.0	3.0
Level 4	3.0	4.0	3.3	2.7	3.0	3.0	3.7	3.0	3.3	3.7
Level 5	4.0	3.7	4.0	4.0	3.0	3.7	4.0	2.3	3.7	3.3
Level 6	4.7	4.7	3.3	3.3	2.0	2.3	3.0	3.7	3.3	4.0
Level 7	3.0	3.7	3.7	3.7	2.3	4.7	4.0	2.7	3.3	2.7
Level 8	2.3	3.0	3.7	4.0	2.7	3.7	3.7	2.7	5.0	2.7
Level 9	3.0	3.7	2.7	3.0	2.7	3.3	3.7	3.7	4.3	3.3
Level 10	3.3	4.7	4.3	3.7	2.3	3.0	4.3	3.3	3.7	3.3
Level 11	4.0	4.7	3.3	3.0	2.7	3.7	3.7	3.3	3.7	3.3
Level 12	3.7	4.0	3.7	2.7	2.7	3.7	3.7	3.3	3.3	3.7
Level 13	5.0	4.7	4.7	4.0	2.3	4.0	4.7	3.3	4.0	4.0
Level 14	4.0	4.0	3.7	3.7	2.7	4.7	4.0	2.7	4.0	3.0
Level 15	2.7	2.7	1.7	1.7	1.3	4.0	4.0	2.3	1.7	3.7

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Resultados de confiabilidad

Como se indicó anteriormente, para medir la confiabilidad de resultados, este trabajo utilizó la fórmula de consistencia interna KR20, la cual funciona de la misma manera que el Alpha Cronbach pero aplicado a formularios con respuestas dicotómicas. En esta instancia, se tomaron los resultados de las pruebas de cada nivel, se les asignó el valor 0 a las respuestas incorrectas y 1 a las respuestas correctas; y luego se le aplicó la fórmula. Los resultados fueron los siguientes:

La figura 11 muestra como los niveles 4 y 7 dieron resultados por encima de 0.9 considerándose *excelente*; por otra parte, las pruebas que correspondieron a los niveles 2 y 13 arrojaron una confiabilidad mayor a 0.8 siendo categorizados como *bueno*; los niveles 3, 6 y 9 más de 0.7 dando como resultado *aceptable;* los correspondientes a los niveles 5, 11 y 15 resultaron mayores a 0.6 y se pueden catalogar como *Cuestionable* y los niveles 12 y 1 resultaron como *Pobre.* Se debe recordar, que los niveles 8, 10 y 14 no formaron parte de la muestra, ya que los profesores de dichos cursos no utilizaron la plataforma Schoology para aplicar la prueba y por lo tanto sus resultados no pertenecen al objeto de estudio.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cuadro 11.

Resultados del KR20		Criterio
Level 01	0.545453222	Pobre
Level 02	0.895119586	Bueno
Level 03	0.756006459	Aceptable
Level 04	0.90318041	Excelente
Level 05	0.697115385	Cuestionable
Level 06	0.786315214	Aceptable
Level 07	0.903186118	Excelente
Level 09	0.736160014	Aceptable
Level 11	0.694864838	Cuestionable
Level 12	0.595628972	Pobre
Level 13	0.800703083	Bueno
Level 15	0.67440225	Cuestionable
Promedio general	0.749011296	Aceptable

Fuente: autoría del investigador

Es muy importante rescatar que ninguna prueba arrojó como resultado menos de 0.5, es decir, ninguna de las pruebas estudiadas se catalogó como *Inaceptable*. Adicionalmente, el promedio general de todas las pruebas es *Aceptable;* dejando espacio importante para mejora.

CONCLUSIONES Y RECOMENDACIONES

Esta investigación buscó establecer una medida de validez para las pruebas de inglés EFL de una institución panameña. Chapelle et al. (2003) explican que el resultado de dicho procedimiento es la construcción de un "argumento de validez". Weir (2005) lo compara con el procedimiento de un abogado defensor que actúa en el tribunal. De esta manera, este abogado debe presentar evidencia que se manifiesta en el argumento de validez y que con base en eso se aprecia que tan confiable es el mismo.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Cada una de las 15 pruebas presenta un caso de validez ligeramente diferente pero dentro de un rango de desempeño similar. A continuación, se presenta un caso de validez para dichas pruebas.

Con respecto a las características de los candidatos:

Las características de los candidatos parecen presentar un nivel de impacto mínimo o irrelevante en los resultados de los estudiantes. La predisposición hacia las pruebas parece ser positiva y existe pocos elementos fisiológicos (dolor de cabeza, resfriado, problema de visión, dislexia, agrafia, entre otros) que causen efecto alguno sobre el desempeño de los candidatos.

También se puede concluir que existe un nivel alto de familiaridad de los estudiantes con el formato y las condiciones del examen, ya que todas las medidas que analizan la predisposición al examen (confianza, preparación, motivación experiencia, entre otros) son positivas.

Con respecto a la validez basada en contexto

Los exámenes cuentan con un nivel poco adecuado de instrucciones y especificaciones que guíen al candidato a lo largo del desarrollo de las pruebas. Tampoco dedican mucha extensión de prueba a explicar los objetivos a lograr o los criterios bajo los cuales serán evaluados.

Uno de los resultados más detectables es la carencia de una evaluación estandarizada oral y auditiva en la prueba. Como se indicó anteriormente, esto no significa que el curso no mide o evalúa el desempeño oral. Lo que sí significa es que dicha evaluación no es estándar, es decir, la aplicación de las evaluaciones orales y auditivas son cambiantes de un curso a otro y de un profesor a otro, permitiendo niveles preocupantes de subjetividad y una posibilidad muy alta de que las calificaciones de los estudiantes en estos aspectos reflejen poco el desempeño real de los candidatos.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

En otro orden de ideas, la mayor parte de los elementos relacionados con las condiciones del examen pueden ser valorados de forma positiva. Los formatos de respuesta, la duración de la prueba, el largo de los textos y los temas contenidos se muestran adecuados, salvo por algunas excepciones pequeñas. Igualmente, las condiciones físicas se muestran consistentes y positivas.

Con respecto a la validez basada en teoría

De la misma manera, existe una valoración adecuada con respecto a los contenidos de las pruebas, teniendo una medición adecuada de habilidades de lectura, resultando mayormente positiva la valoración de las pruebas en su capacidad de medir habilidades de lectura superficial y encontrar información específica en el texto.

En el mismo orden de ideas, las pruebas demostraron, en su mayoría, poder medir de forma satisfactoria las habilidades relacionadas con dominio de gramática de los candidatos, comprobando exitosamente la capacidad de los candidatos de entender cohesión y coherencia de mensajes, así como también su habilidad de interpretar significado y entender el entorno sociolingüístico.

Con respecto a la confiabilidad de resultados o validez de calificación

Este el elemento menos consistente entre las diferentes pruebas, ya que, al aplicar la fórmula KR20 a los resultados variaron dramáticamente entre *Cuestionable* y *Excelente.*

También se debe tomar en cuenta que la presente investigación solamente midió la confiabilidad de consistencia interna, la cual es sólo uno de los métodos de confiabilidad de resultados sugeridos por AERA/APA/NCME (1974, 1985, 1999). Es probable que la aplicación de más métodos de confiabilidad de resultados hubiera arrojado resultados más precisos.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

Sin embargo, partiendo de los resultados obtenidos, se puede inferir que existe un espacio muy amplio de mejora en el contenido de las pruebas para que éstas tengan mejor consistencia interna en sus aplicaciones.

El argumento general de validez para los instrumentos de evaluación del curso de inglés EFL de la institución panameña estudiada se puede considerar *Aceptable,* pero con un espacio muy amplio de crecimiento y mejora si son aplicados ciertos cambios y políticas que incrementarían su medida de validez. A continuación se presentan recomendaciones que podrían ayudar a alcanzar dicho crecimiento.

Recomendaciones

- Agregar a las pruebas una sección explicativa que muestre a los candidatos los objetivos de la prueba y los criterios de evaluación y ponderación de calificaciones.
- Crear pruebas orales estándar, basadas en los trabajos de evaluación oral de Weir (2005) y O'Sullivan etal. (2002), que contengan rubricas de evaluación y que comprueben un listado extenso de competencias en cada candidato, acompañado de entrenamiento constante a los evaluadores para disminuir los elementos de subjetividad.
- Agregar una sección obligatoria de evaluación auditiva con ítems que evalúen la capacidad de comprensión de información general, así como la comprensión de información específica.
- Llevar a cabo un estudio más profundo de confiabilidad de resultados, utilizando los métodos sugeridos por AERA/APA/NCME (1999): Confiabilidad examinación-reexaminación; confiabilidad de formularios y confiabilidad de resultados.
- Realizar nuevamente la prueba de consistencia interna KR20 con un período mayor y una cantidad más consistente de aplicaciones.
-

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

- Llevar adelante esfuerzos para fomentar el uso de pruebas digitales en vez de pruebas en papel para tener un mejor registro.
- Aplicar todas las recomendaciones anteriormente mencionadas y en un período de 6 meses a un año realizar nuevamente un proceso de validación, así como también, establecer una agenda de validación de pruebas anual.

QLU QUALITY LEADERSHIP UNIVERSITY ISSN-L 2644-4038

REFERENCIAS

Alderson, J.C. (1991). Dis-sporting Life. Response to Alistair Pollit's paper. In Alderson and North

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1974, 1985, 1999). Standards for Educational and Psychological Testing. Washington, DC: Author.

Anastasi, A. (1988). Psychological Testing (6th edition). New York: Macmillan.

Bachman, L.F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L. and Palmer, A. (1996). Language Testing in Practice. Oxford: Oxford University Press.

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

Chapelle C., Jamieson J. y Hegelheimer V. (2003); Validation of a web-based EFL test

Kokhan K.(2013); An argument against using standardized test scores for placement of international undergraduate students in (ESL) courses.

Kuder, G.F. y Richardson, M.W. Psychometrika (1937) The theory of the estimation of test reliability 2: 151. https://doi-org.echo.louisville.edu/10.1007/BF02288391

O'Sullivan, B. (2000). Towards a Model of Performance in Oral Language Testing. Unpublished PhD Dissertation. University of Reading.

Pennington, D. (2003). Essential Personality. Arnold. p. 37. ISBN 0-340-76118-0.

Weir, C. (2005); Language Testing and Validation: An Evidence-Based Approach. Springer