ANÁLISIS DEL FUNCIONAMIENTO DIFERENCIAL DE LOS REACTIVOS DE LA PRUEBA DE HABILIDADES VERBALES

2021

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

SECRETARÍA GENERAL

COORDINACIÓN DE UNIVERSIDAD ABIERTA, INNOVACIÓN EDUCATIVA Y EDUCACIÓN A DISTANCIA

CONSEJO DE EVALUACIÓN EDUCATIVA

COMISIÓN PERMANENTE DEL POSGRADO

Comisión Permanente del Posgrado, Red Colaborativa 1

Dra. Gloria Vilaclara Fatjó
Dra. Amelia Farrés González Saravia
Dr. José Ramón Hernández Santana
Mtra. Ana María Losada Alfaro
Dra. Maricela Luna Muñoz
Dra. Marisa Mazari Hiriart
Dr. Alfonso Gerardo Navarro Sigüenza
Dra. Aurea Orozco Rivas
Dr. Enrique Pérez Campuzano
Dr. Héctor Quiroz Rothe
Dr. Carlos Humberto Reyes Díaz
Dra. Jeanett Reynoso Noverón
Lic. José Javier Saldaña Solís
Dra. Christina Siebe Grabach
Dra. Elizabeth Solleiro Rebolledo
Dra. Abril Uscanga Barradas
Dr. Erik Velásquez García

Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia
Dirección de Evaluación Educativa

Coordinador
Dr. Melchor Sánchez Mendiola

Director de Evaluación Educativa
Dr. Adrián Martínez González

Subdirectora de Evaluación de Posgrado y Titulación
Mtra. Nancy Sofía Contreras Michel

Coordinador de Análisis de Resultados de Evaluación Educativa
Mtro. Manuel García Minjares

Coordinación técnico-académica

Mtra. María Juliana Londoño Cárdenas
Ana Itzel Pascual Vigil
Con el apoyo de
Lic. Davin Eduardo Díaz García

Tabla de contenido

Introducción 1

Procesos de admisión al Posgrado 2

Prueba diagnóstica de habilidades verbales para el proceso de admisión de la UNAM 4

Propósito del estudio 7

Método 8

Participantes 8

Prueba de Habilidades Verbales 9

Recolección de datos 11

Análisis del Funcionamiento Diferencial de Reactivos (DIF) 11

a) Método de Mantel-Haenszel 12

b) Método de diferencia de logits con modelo de Rasch 12

Resultados 14

Comprensión lectora 14

Redacción y gramática 17

Conclusiones 20

Recomendaciones 22

Referencias 23

Anexos 24

Anexo A. Tabla de especificaciones de Comprensión lectora

Anexo B. Tabla de especificaciones de Redacción y gramática

Introducción

En 2017 la Red Colaborativa 1 de la Comisión Permanente del Posgrado del Consejo de Evaluación Educativa de la Universidad Nacional Autónoma de México (UNAM) concluyó que la evaluación de habilidades generales en los aspirantes era necesaria para una gran mayoría de los programas de posgrado. Los componentes de esta prueba deberían ser los ejes transversales para una evaluación complementaria de ingreso a los programas de posgrado en las cuatro Áreas de Conocimiento.

Con base en una búsqueda sobre otros exámenes de admisión (e.g., GRE y EXANI III) se propuso desarrollar las evaluaciones en: comprensión de textos, redacción y gramática, habilidades digitales y pensamiento crítico. Los Consejeros de este Comisión consideraron pertinente realizar un estudio que valorara la eficacia de la Prueba de Habilidades Verbales, diseñada por la Dirección de Evaluación Educativa (DEE) de la Coordinación de Universidad Abierta Innovación Educativa y Educación a Distancia (CUAIEED) como un instrumento para ser aplicado a los aspirantes a ingresar a alguno de los programas de posgrado independientemente del área de conocimiento.

En este informe se describen los resultados sobre el análisis del funcionamiento diferencial de los reactivos de la Prueba de Habilidades Verbales. El informe se encuentra organizado en cinco apartados:

En el apartado de *Procesos de admisión al posgrado* se presenta brevemente el uso de pruebas objetivas en los procesos de selección y la necesidad de incluir una evaluación de las habilidades verbales a los aspirantes al posgrado. En *Método* se describen los participantes, el diseño y aplicación de la prueba, así como el proceso de análisis estadístico de los reactivos para detectar un comportamiento diferencial (DIF). Luego en la sección de *Resultados* se presentan los hallazgos de los análisis de detección de DIF; en *Conclusiones*, se sintetizan los datos más importantes y en *Recomendaciones* se presentan sugerencias para mejorar la elaboración de futuras pruebas e instrumentos.

Procesos de admisión al Posgrado

Las Instituciones de Educación Superior (IES) se enfrentan a una serie de retos importantes para cumplir con una enseñanza de calidad. Particularmente en los procesos de admisión cuestiones como imparcialidad, equidad y oportunidades de aprendizaje son puntos clave a considerar. Especialmente cuando hay una gran demanda y un número limitado de lugares disponibles para el ingreso al posgrado. Uno de los enfoques utilizados en estos procesos de admisión es el que se basa en los requisitos y méritos académicos. Este enfoque tiene implicaciones importantes con respecto a la validez y a la imparcialidad de los procedimientos de selectividad que se aplican en las universidades (Oliveri y Wendler, 2020).

Entre estas implicaciones se destaca la validez del uso e interpretación de resultados obtenidos en la aplicación de instrumentos de medición que son utilizados como criterio de admisión. Una de las interrogantes fundamentales es la capacidad predictiva de las evaluaciones realizadas con respecto al éxito académico de los examinados en los programas para los que sirven como filtro de admisión. En otras palabras, es crucial presentar evidencia de que las evaluaciones utilizadas realmente sirven como un filtro que permite seleccionar a aquellos con mayor probabilidad de culminar con éxito los respectivos programas de estudios. En ese sentido, se espera que los instrumentos empleados sean, no solamente pertinentes en cuanto a sus contenidos, sino también sensibles ante las poblaciones de aspirantes, de manera que los instrumentos logren un nivel de imparcialidad que les permita evaluar el nivel de habilidad o atributo que se pretende medir, y no verse sesgados por otras variables o aspectos particulares de las distintas poblaciones sustentantes.

En el caso de los programas de posgrado de la UNAM los procesos de admisión responden a las necesidades particulares de los diversos campos de conocimiento, de manera que se diseñan y aplican instrumentos de evaluación específicos a las cuatro áreas de conocimiento: I: Ciencias Físico-Matemáticas y de las Ingenierías (CFMI); II: Ciencias Biológicas y de la Salud (CBQS); III: Ciencias Sociales (CS) y IV:

Humanidades y de las Artes. Con el interés de generar una herramienta para el proceso de selección que permitiera detectar habilidades esenciales para el buen desempeño dentro del posgrado se propone la construcción de la prueba de Habilidades Verbales con la posibilidad de integrarla como una herramienta de evaluación general que pudiera aplicarse en todos los programas de posgrado y áreas de conocimiento por igual. Este uso previsto para la prueba de Habilidades verbales implica retos importantes con respecto a la imparcialidad de los resultados de los sustentantes de cada una de las áreas de conocimiento.

Imparcialidad

El concepto de imparcialidad (*fairness*) en el contexto de diseño, elaboración, aplicación uso de instrumentos en psicología y educación ha sido ampliamente estudiado y ha cobrado mayor relevancia en los últimos años. El concepto es bastante amplio y se utiliza para abordar distintas etapas y procesos en la creación de pruebas. Por ejemplo, en los Estándares para pruebas educativas y psicológicas (AERA, 2014), se reconoce que estudiar la imparcialidad podría implicar abordar desigualdades sociales y muchos otros aspectos no directamente relacionados con la prueba. En términos generales, en los Estándares se define imparcialidad como "la capacidad de respuesta a características individuales y contextos de evaluación de modo que los puntajes de la prueba arrojen interpretaciones válidas para los usos previstos" (AERA, 2014). Se destaca dentro de las áreas de mayor importancia la referente al sesgo a nivel de medición.

Sesgo estadístico

Dentro de los Estándares, se reconoce al sesgo de medición como "…una amenaza central a la imparcialidad de la prueba" (AERA, 2014). En términos generales, el sesgo de medición se presenta cuando una prueba favorece a grupos específicos por encima de otros. De acuerdo con Bond (1996), podemos afirmar que el sesgo se refiere a

> "… la medida en que la puntuación y el uso de una prueba son válidos para todos los individuos y grupos previstos. Como tal, si una evaluación da como resultado puntuaciones que subestiman

> sistemáticamente el estado de los miembros de un grupo en particular en el constructo en cuestión, entonces la prueba está sesgada en contra de los miembros de ese grupo". (pág.119)

Esta definición implica que, para detectar sesgos, es necesario contar un estándar de referencia que nos permita identificar el nivel de atributo real de los distintos grupos, para con ello poder identificar si en realidad la prueba que estamos utilizando presenta diferencias importantes entre un grupo y otro en referencia a su nivel real de atributo (Pendfield & Camilli, 2006). Dada la complejidad de generar este contraste, una alternativa por la que se ha optado es por analizar el sesgo de medición a nivel de los reactivos.

Importancia de DIF

En este intento de analizar la posible presencia de sesgos a nivel de reactivo, se generó un marco de referencia estadístico que hoy en día es denominado como Funcionamiento Diferencial del Reactivo (DIF, por sus siglas en inglés). El término fue acuñado por primera vez por Holland y Thayer (1988). En términos generales, podemos decir que existe DIF "… cuando examinados con iguales capacidades difieren en sus probabilidades de responder a un reactivo de la prueba correctamente como una función de pertenencia a un grupo" (Holland y Thayer, 1988). El análisis DIF tiene gran importancia en este contexto, ya que es un método que busca fomentar imparcialidad en los resultados de la prueba para garantizar que ningún grupo o área específica parta con ventaja al responder el instrumento.

Prueba diagnóstica de habilidades verbales para el proceso de admisión de la UNAM

Las habilidades del manejo del lenguaje representan una herramienta fundamental para el aprendizaje y el desarrollo de las personas dado que son necesarias para la construcción y comunicación del conocimiento. En la Educación Superior, se exige el contacto permanente con habilidades del lenguaje como lo son la lectura y la escritura académica. Esto hace que las actividades como leer y escribir en la universidad se tornen complejas, ya que se requiere una formación de nivel avanzado que le permita al

estudiante no solamente asimilar la información que lee, sino también procurar la generación escrita de procesos de investigación y resolución de problemas (Cisneros, M., Olave, G. y Rojas, I.,2013).

Bajo este panorama, algunos de los integrantes de diferentes programas de posgrado de la UNAM reflexionaron sobre el papel que tienen estas habilidades en las experiencias que viven día con día en los espacios educativos. Esta reflexión los llevó a la conclusión de que existen deficiencias heredadas de niveles anteriores que dificultan su quehacer docente y la graduación de los estudiantes. Por esta razón, algunos de los programas de posgrado de la UNAM consideraron que la Prueba de Habilidades Verbales, construida por la DEE, es un instrumento fundamental para la selección de sus aspirantes.

La prueba tiene como finalidad distinguir a los aspirantes que cuentan con las habilidades de comprensión lectora y del uso apropiado de la gramática. Ambas han sido consideradas por los posgrados como necesarias para el cumplimiento exitoso de los estudios de cada programa, ya que fomentan la graduación en los tiempos curriculares adecuados y con ello se ve favorecida la eficiencia terminal. Cabe destacar que existen varias universidades a lo largo del mundo que utilizan exámenes de comprensión lectora y otras habilidades externas a los campos de conocimiento de los programas como un medio de selección de candidatos, encontrando evidencia importante del valor de dichas habilidades (Oliveri y Wendler, 2020).

Esta prueba fue implementada con anterioridad en diferentes áreas de conocimiento de la Universidad obteniendo resultados similares en relación al desempeño de los aspirantes. Para comprobar su valor como un instrumento que apoya a los posgrados, en el año 2017 la DEE generó un estudio del valor predictivo del examen de ingreso a un Posgrado del Área de las Ciencias Sociales sobre la relación entre los resultados de los aspirantes en componentes que evalúan sus habilidades verbales y su desempeño en el posgrado. En este estudio se analizó, a través una regresión lineal múltiple, la contribución que tiene la edad, el género, los antecedentes escolares y el puntaje en el examen de selección sobre el desempeño académico. Los modelos generados se usaron para predecir las calificaciones de los alumnos durante el año que cursan la especialización y que presentaron la prueba en las generaciones 2015 y 2016.

En el modelo presentado en la Tabla 1 se considera que los resultados obtenidos en el componente Redacción y gramática del examen de ingreso son predictivos del desempeño del segundo semestre de las especializaciones con una presencia mayor de métodos cuantitativos en su plan de estudios. Esto muestra que, por cada punto porcentual del porcentaje de aciertos obtenido por un aspirante en este componente, su promedio de calificaciones en el segundo semestre incrementará 0.45.

Tabla *1*
Modelo para especializaciones con mayor presencia de métodos cuantitativos. Generación 2015

		B	E.E	b^+	R^2	ΔR^2	P
		Segundo semestre					
Modelo 2	Redacción y gramática	0.200	0.08	0.43	0.18		0.017

$p<0.05$
Fuente: DEE (2017)

Estos resultados sugieren que la evaluación de las habilidades verbales de los aspirantes a los posgrados puede aportar información valiosa sobre su desempeño posterior en los planes de estudio, lo cual sería de utilidad para los alumnos, tutores y coordinadores de posgrado en la definición de estrategias para incentivar la mejora de estas habilidades y la importancia de su evaluación en los procesos de selección.

Propósito del estudio

El presente informe de evaluación pretende abordar tres propósitos principales:

- Generar evidencias de validez de la interpretación de que los resultados de la prueba de Habilidades verbales son imparciales. Esto mediante el uso de técnicas para detectar DIF.
- Comparar distintas metodologías para la obtención del efecto DIF en los reactivos de la prueba para identificar los métodos más adecuados, precisos y eficaces en la detección de DIF en este contexto.
- Reportar aquellos reactivos que presenten DIF directamente a los interesados, con la finalidad de que puedan tomar las medidas necesarias a nivel de corrección, ponderación o eliminación de los reactivos para la calificación final de los sustentantes, reduciendo así la presencia de sesgos y fomentando la imparcialidad dentro del proceso de selección.

Método

En esta sección se describen los componentes del método del estudio: los participantes, los instrumentos diseñados para la recolección de la información, el proceso de recolección de la información, así como los criterios que se adoptaron para el análisis de los datos.

Participantes

Con la finalidad de contar con una muestra proveniente de las cuatro áreas de conocimiento del Posgrado de la UNAM, se aplicó el instrumento en los procesos de admisión de siete programas de posgrado en diferentes convocatorias, desde el ciclo escolar 2019-1 al 2020-2. La muestra total analizada para este informe fue 1218 sustentantes, con 147 del Área CFMI, 417 del Área CBQS, 427 del Área CS y 227 del Área HA (Tabla 2). Todos los posgrados participantes de esta aplicación participaron de manera voluntaria.

Tabla *2*
Número de posgrados y examinados por área de conocimiento y periodo de aplicación

Área	Posgrados participantes	Periodo(s) de aplicación	Examinados
I. Ciencias Físico - Matemáticas y de las Ingenierías (CFMI)	1	2019-2	147
II. Ciencias Biológicas, Químicas y de la Salud (CBQS)	2	2019-1 2019-2	417
III. Ciencias Sociales (CS)	2	2019-2 2020-1	427
IV. Humanidades y de las Artes (HA)	2	2020-1 2020-2	227

Prueba de Habilidades Verbales

Para la construcción y diseño del instrumento, una comisión de académicos expertos en español elaboró las tablas de especificaciones de los dos componentes que comprende la Prueba: *Comprensión lectora* y *Redacción y gramática*.

Las tablas de especificaciones se integraron con los temas y subtemas que la comisión de expertos determinó como fundamentales, así como los correspondientes resultados de aprendizaje que se desean evaluar, su nivel cognoscitivo y el peso específico. Con base en estas tablas y con la asesoría de la DEE, los profesores elaboraron y validaron los reactivos de la prueba.

Para ensamblar la prueba solo se consideraron los reactivos con indicadores estadísticos deseables de acuerdo con la Teoría Clásica de los Tests (TCT) y los modelos de uno y dos parámetros de la Teoría de Respuesta al Ítem (TRI). Cabe señalar que estos reactivos fueron aplicados a diferentes poblaciones de aspirantes a programas de posgrados de la UNAM, y posteriormente se calibraron, a fin de contar con sus indicadores estadísticos.

Para la selección de los reactivos se analizaron los índices de dificultad, de discriminación y los coeficientes de correlación biserial del reactivo y de cada una de sus opciones de respuesta establecidos en la TCT. En lo que respecta al índice de dificultad se incluyeron en la prueba reactivos con una dificultad dentro del rango de 0.20 a 0.80. En el índice de discriminación se consideraron reactivos que permitieran distinguir entre los grupos de alto y bajo desempeño[1]. Se consideraron reactivos aceptables aquellos que contaran con índices de discriminación con valores iguales o mayores a 0.20. Un reactivo discrimina de manera eficaz si lo responden correctamente más examinados con una puntuación alta en la prueba. También se cuidó el comportamiento de las opciones de respuesta, es decir, se seleccionaron reactivos donde la

[1] Generalmente se considera como como grupo de alto desempeño al 27% de la muestra de examinados que obtuvieron el mayor número de aciertos y como de bajo desempeño al 27% con menor número de aciertos.

opción correcta fuera elegida por una mayor proporción del grupo de alto desempeño en comparación con el de bajo desempeño; en las opciones incorrectas se verificó que las eligiera una mayor proporción del grupo de bajo desempeño. En el coeficiente de correlación punto biserial se buscó que los reactivos tuvieran un valor igual o mayor a 0.20. Este valor positivo indica que los examinados que contestaron correctamente el reactivo obtuvieron un mayor número de aciertos en la prueba.

En el caso de la TRI se empleó el modelo de Rasch que toma en cuenta la dificultad del reactivo. En este parámetro se consideraron reactivos que tuvieran valores de dificultad entre –2.5 y +2.5. También se consideró el modelo de dos parámetros, en donde, la discriminación de los reactivos elegidos tenía un valor mayor a 0.45.

Finalmente, la prueba se conformó con 31 reactivos de opción múltiple con cuatro opciones de respuesta que miden *Comprensión lectora* y *Redacción y gramática*, que cumplieron los requisitos anteriormente mencionados, (ver Tabla 3).

Tabla *3*
Estructura de la Prueba de Habilidades Verbales

Componentes	Reactivos	Porcentaje
Comprensión lectora	11	35.5
Redacción y gramática	20	64.5
Total	**31**	**100.0**

Se diseñaron dos versiones de la prueba integradas con los mismos reactivos, las cuales sólo difieren en el orden en el que se encuentran presentados los componentes, (ver Tabla 4).

Tabla *4*
Distribución de los componentes en las versiones A y B

Versión A	Versión B
Comprensión lectora	Redacción y gramática
Redacción y gramática	Comprensión lectora

Recolección de datos

La DEE planeó la logística de la aplicación lápiz-papel y en línea con supervisión presencial para la prueba de Habilidad Verbal en siete programas de posgrado. Esto implica personalizar el material para los examinados e integrar los formatos necesarios para los aplicadores, a quienes se capacitó con los lineamientos establecidos para la aplicación de una prueba objetiva.

La aplicación de la prueba estuvo a cargo del personal de los programas de posgrado y fue supervisada por personal de la DEE, quienes recibieron y resguardaron el material utilizado. Posteriormente, en el caso de las pruebas en formato lápiz-papel, se leyeron las hojas de respuesta en un lector óptico para obtener la cadena de respuestas de los examinados. En las aplicaciones en línea, la Unidad de Sistemas para la Evaluación Educativa de la DEE extrajo del sistema de aplicación de pruebas, las cadenas de respuestas. En ambos casos, se verificó que las bases de respuestas estuvieran libres de elementos que dificultaran su procesamiento y fueron entregadas a la Coordinación de Análisis de Resultados de Evaluación Educativa para que los reactivos fueran calibrados a partir de los índices de la TCT y los modelos de uno y dos parámetros de la TRI.

Análisis del Funcionamiento Diferencial de Reactivos (DIF)

El análisis estadístico de los reactivos de la prueba se realizó por medio del análisis del Funcionamiento Diferencial del Reactivo (DIF, por sus siglas en inglés). El análisis DIF permite identificar la presencia de sesgos en los reactivos que favorezcan más a algún grupo con respecto a otro independientemente de su nivel de habilidad. Esto es, si dos examinados tienen el mismo nivel de habilidad, pero uno de ellos por pertenecer a un grupo distinto tiene mayor probabilidad de contestar de manera correcta dicho reactivo. Existen muchas y muy diversas metodologías para detectar la presencia de DIF, algunas de ellas directamente derivadas de la TRI, y algunas otras de los principios y supuestos de la TCT.

Para el presente informe, se utilizaron dos metodologías principales para realizar el análisis DIF comparando las cuatro áreas de conocimiento de los posgrados UNAM. Para poder obtener las diferencias específicas que se presentan en la comparativa de un área y otra directamente, se siguieron las recomendaciones de la Educational Testing Service (ETS), utilizando el método de Mantel-Haenszel y el método de diferencias de logits basado en un modelo Rasch (Zwick, 2012). A continuación, se describen brevemente ambos métodos y la interpretación de sus resultados.

a) Método de Mantel-Haenszel

La prueba de Mantel-Haenszel (MH) es un método de estimación que no se basa en la TRI, y que suele usarse principalmente para comparación de dos grupos; busca probar si existe una asociación entre la pertenencia a cierto grupo y la respuesta correcta a un reactivo, tomando como condición el puntaje total de la prueba.

El valor que se obtiene de calcular este estadístico sigue una distribución χ^2 con un grado de libertad. Por lo tanto, valores de la prueba MH mayores a un valor crítico basado en la distribución χ^2, nos indicará la presencia de DIF para ese reactivo en específico. Este valor suele transformarse en la razón de probabilidades logarítmica. Bajo esta transformación, se clasifica el tamaño del efecto DIF como insignificante si es menor a 1, moderado si va de 1 a 1.5 y grande si es mayor a 1.5 (Zwick, 2012).

b) Método de diferencia de logits con modelo de Rasch

Dado que hablar de DIF implica hablar de diferencias en probabilidades de respuesta, la manera más lógica y comprensible de interpretarlo es a través de la TRI. Esta teoría parte del supuesto de que la probabilidad de responder de manera correcta un reactivo está dada en función del nivel de habilidad de los sustentantes y por uno o más parámetros referentes al reactivo de manera específica. Así, al utilizar el modelo de Rasch, una vez que se obtiene el nivel de habilidad de todos los participantes, se puede calcular el parámetro de dificultad para los dos grupos por separado y observar si existen diferencias. El criterio

que se tomó como punto de corte para detectar la presencia de DIF fue una diferencia en lógitos mayor a 0.43 y para clasificar el nivel de DIF se utilizaron los criterios determinados por el ETS (Zwick, 2012). En el caso que la diferencia entre los dos grupos comparados no supere los 0.43 lógitos se considera un DIF sin importancia; si es mayor a 0.64 lógitos indica DIF moderado-alto y si la diferencia se encuentra entre estos valores se trata de un DIF leve-moderado (Holland y Weiner, 1993).

Software utilizado

Los análisis estadísticos fueron elaborados mediante el *software R* en su versión 4.1.0. Para los análisis DIF, se utilizaron los paquetes difR (Magi, et al., 2020) para el método Mantel-Haenszel y para la realización del análisis de Rasch se usaron el eRm y BILOG-MG.

Resultados

Los resultados de los análisis del funcionamiento diferencial de los reactivos por área de conocimiento del posgrado están organizados por cada componente de la *Prueba de Habilidades Verbales.* Los métodos de detección del DIF utilizados en este estudio realizan una comparación entre dos grupos de poblaciones. Por lo tanto, los análisis DIF por el método de Mantel-Haenszel y el modelo de Rasch se llevaron a cabo considerando seis arreglos de comparación de áreas de conocimiento del posgrado: CFMI y CBQS; CFMI y CS; CFMI vs HA; CBQS vs CS; CBQS vs HA y CS vs HA.

Comprensión lectora

Método de Mantel-Haenszel

El análisis DIF mediante el método de Mantel-Haenszel se realizó por pares de áreas de conocimiento del posgrado. En la tabla 5 se puede observar de manera sintética los resultados de dicho análisis. Se presentan en cada columna los resultados de cada análisis por pares, mostrando el valor de DeltaMH (D-MH) y su interpretación mediante códigos de signos positivos y negativos. Los valores positivos indican que el reactivo de *Comprensión Lectora* (CL) resulta más fácil al primero de los dos grupos de cada comparación, mientras que los negativos indican que resulta más fácil para el segundo. La presencia de doble signo (i.e. “++” o “--”) indica que el funcionamiento diferencial del reactivo es elevado; cuando es únicamente un solo signo (i.e. “+” o “-”), indica que el DIF es moderado, mientras que los reactivos que no presentan ningún signo indican que el DIF es leve o insignificante.

Tabla *5*

Presencia de DIF en reactivos de comprensión lectora, por áreas de conocimiento (Método MH)

	CFMI vs CBQS		*CFMI vs CS*		*CFMI vs HA*		*CBQS vs CS*		*CBQS vs HA*		*CS vs HA*	
	D-MH	*DIF*	*D-MH*	*DIF*	*D-MH*	*DIF*	*D-MH*	*DIF*	*D-MH*	*DIF*	*D-MH*	*DIF*
CL1	-0.3413		-0.0772		-0.6019		0.3590		0.2636		-0.5766	
CL2	-0.1132		0.2011		0.6895		0.5407		1.0392	+	0.2054	
CL3	0.3135		0.2910		-0.1509		0.1748		0.0082		-0.5074	
CL4	0.6346		1.6630	++	1.8517	++	1.4058	+	1.8103	++	0.1927	
CL5	-0.5681		-0.9614		-1.0373	-	-0.2590		0.2573		-0.3296	
CL6	-0.2327		0.6280		1.5383	++	1.0111	+	2.3706	++	1.0959	+
CL7	-0.2605		-0.8538		-0.7461		-0.3309		-0.2827		-0.0347	
CL8	0.3600		0.6940		1.3485	+	0.4958		1.3581	+	0.3526	
CL9	0.8881		-0.0464		0.7109		-0.6806		-0.1651		0.3982	
CL10	-1.1381	-	0.4230		0.9538		1.4919	+	2.0537	++	0.1688	
CL11	0.4672		-0.2909		0.2710		-0.4740		-0.0945		0.1402	

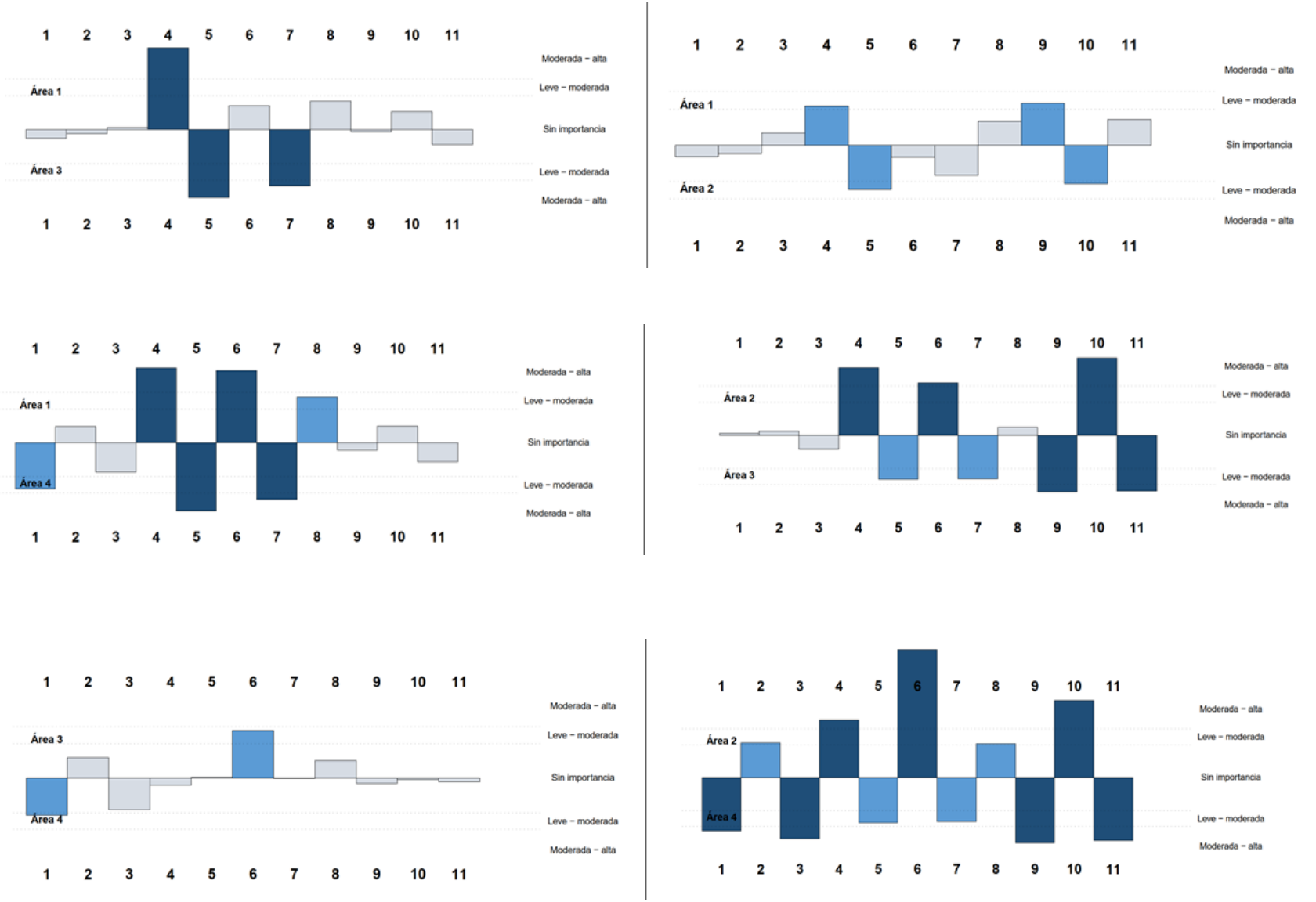
Dentro de este análisis, 6 de los 11 reactivos presentan DIF en al menos una de las comparativas (CL2, CL4, CL5, CL6, CL8 y CL10). Se puede observar que en la comparativa directa, los reactivos con mayor presencia de DIF significativo son CL4, CL6 y CL10, siendo el reactivo 4 el que presenta mayor nivel de DIF, especialmente al comparar el área CFMI con las áreas CS y HA. Cabe destacar que los reactivos 1, 3, 7, 9 y 11 no presentan DIF significativo en ninguna de las comparativas. Por otro lado, las comparativas entre CFMI y CBQS y CS con HA son las que menos reactivos con DIF presentan.

Método de diferencia de logits con modelo de Rasch

En este análisis de DIF se obtiene un coeficiente que representa la diferencia en términos del parámetro de dificultad en el modelo de Rasch. Se utilizó como criterio de detección de DIF una diferencia mayor a 0.5 en dicho coeficiente siendo los valores positivos índices de que el reactivo es más difícil para el primer grupo de la comparación, y los valores negativos indicadores de que es más difícil para el segundo grupo. En la Figura 1 se representa, por medio del color de la barra, el nivel de DIF detectado en los reactivos de este componente. El azul más oscuro indica DIF alto, el azul claro DIF moderado y el gris DIF leve.

Figura 1.

Presencia de DIF en reactivos de Comprensión lectora, por áreas de conocimiento (Método Rasch)



En este análisis se identificaron nueve reactivos con DIF moderado alto (CL1, CL3, CL4, CL5, CL6, CL7, CL9, CL10 y CL11). En este caso, se puede observar que los resultados son similares a los obtenidos con el método Mantel-Haenszel, siendo que los principales reactivos detectados con DIF son el 4, 6 y 10, y que la comparativa entre CFMI (Área 1) y CBQS (Área 2), así como la comparativa entre CS (Área 3) y HA (Área 4) no presentan reactivos con DIF significativo.

Redacción y gramática

Método de Mantel-Haenszel

En la Tabla 6 se puede observar la presencia de DIF en al menos una de las comparativas en 13 de los 20 reactivos (RG2, RG3, RG4, RG5, RG6, RG7, RG9, RG10, RG11, RG14, RG17, RG18 y RG20). En este caso los reactivos con mayor presencia de DIF a través de distintas comparaciones son RG2, RG5, RG10 y RG14, siendo los únicos que presentan DIF en tres o más de las comparativas. De igual forma como sucedió en *Comprensión lectora*, las comparativas con menor presencia de reactivos con DIF son la comparación entre CFMI y CBQS y entre CS y HA.

Método de diferencia de logits con modelo de Rasch

Los resultados del análisis bajo el modelo de Rasch se presentan en la Figura 2 con todas las comparaciones entre áreas y con el criterio de diferencia en el parámetro de dificultad. En al menos una de las comparativas 13 reactivos presentaron DIF moderado alto en *Redacción y Gramática*: RG2, RG3, RG4, RG6, RG7, RG9, RG10, RG11, RG14, RG17, RG18, RG19 y RG20. Nuevamente los resultados son consistentes con los observados en el análisis con método Mantel-Haenszel, siendo los reactivos que presentan DIF en el mayor número de comparaciones los reactivos 2, 6, 10, 11, 14 y 17. Nuevamente, las comparativas entre FMI y CBQS y entre CS y HA son las que presentan menos DIF.

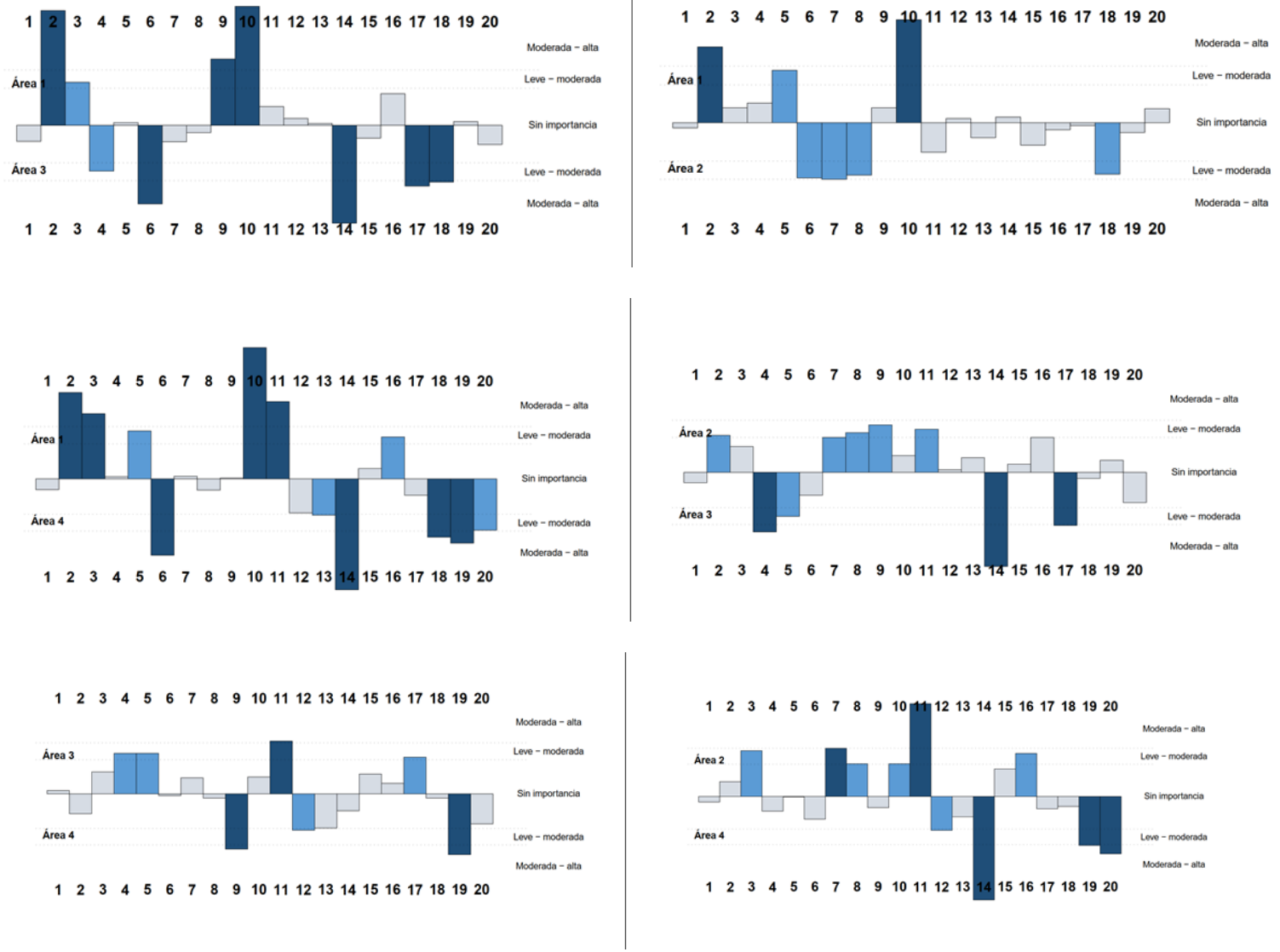
Tabla *6*

Presencia de DIF en reactivos de Redacción y gramática, por áreas de conocimiento (Método HM)

	CFMI vs CBQS		CFMI vs CS		CFMI vs HA		CBQS vs CS		CBQS vs HA		CS vs HA	
	D-MH	DIF	D-MH	DIF	D-MH	DIF	D-MH	DIF	D-MH	DIF	D-MH	DIF
RG1	0.2110		-0.2285		0.0593		-0.5118		-0.258		0.3048	
RG2	1.5568	++	2.2044	++	2.1050	++	0.286007		0.4596		-0.5762	
RG3	0.5464		0.9938		2.2370	++	0.0176		1.2828	-	0.8303	
RG4	0.6562		-0.7378		0.1260		-1.6812	--	-0.3744		0.8234	
RG5	1.2258	+	-0.0113		1.5244	++	-1.4521	-	0.1320		1.0008	+
RG6	-0.7830		-1.3971	-	-1.3461	-	-0.9064		-0.5914		0.0377	
RG7	-0.8015		-0.2243		0.4457		0.2552		1.1173	-	0.2498	
RG8	-0.7232		-0.1787		0.0960		0.5384		0.5102		-0.0343	
RG9	0.4842		1.4956	+	0.5482		0.5213		-0.2519		-1.1407	-
RG10	1.9686	++	2.3729	++	2.4915	++	0.0218		0.3846		0.0279	
RG11	-0.2629		0.4735		1.1990	+	0.5332		1.5132	++	0.8711	
RG12	0.1972		0.2789		-0.0854		-0.6056		-0.6401		-0.4567	
RG13	-0.0629		0.1092		-0.8703		-0.1726		-0.5686		-0.6968	
RG14	0.3519		-1.6112	--	-1.8522	--	-2.4296	--	-2.3678	--	-0.3806	
RG15	-0.2352		-0.2516		0.3636		-0.227		0.3420		0.1126	
RG16	0.1656		0.6961		0.8200		0.3393		0.7761		-0.0746	
RG17	0.1574		-1.0964	-	-0.0557		-1.5332	--	-0.3126		0.9690	
RG18	-0.7694		-1.0315	-	-0.9224		-0.5537		-0.1869		-0.0015	
RG19	0.0039		0.1744		-0.8011		-0.3387		-0.9962		-0.9895	
RG20	0.5440		-0.3638		-0.4330		-1.1748	-	-1.1294	-	-0.3004	

Figura 2.

Presencia de DIF en reactivos de Redacción y gramática, por áreas de conocimiento (Método Rasch)



Conclusiones

De manera general, se observan resultados consistentes en cuanto a los reactivos con mayor presencia de DIF a través de los distintos análisis realizados. Entre los resultados principales, se destacan las diferencias observadas entre las distintas áreas de conocimiento, siendo que los reactivos de los componentes de la prueba funcionaban de manera similar entre las áreas CS y HA, y en menor medida entre las áreas CBQS y FMI, mientras que, en las comparaciones en el resto de los pares, se observó mayor funcionamiento diferencial en los reactivos.

Pese a que una proporción importante de reactivos de cada componente presenta funcionamiento diferencial, es importante notar que la mayoría de ellos presentan un funcionamiento diferencial moderado, especialmente tomando en cuenta el análisis con el método de Mantel-Haenszel.

De igual forma, es importante destacar que, la presencia de DIF no es un indicador contundente de que existan problemas específicos referentes al reactivo; lo que la presencia de DIF nos señala es que un reactivo funciona de manera distinta en dos poblaciones, independientemente del nivel de habilidad de las personas. Por ello, si bien la presencia de DIF puede ser asociada a deficiencias en los reactivos, también puede ser indicio de diferencias reales entre ambas poblaciones. El contenido de un reactivo puede ser teóricamente adecuado y no presentar sesgos específicos que favorezcan a alguna población sobre otra, y aun así, pueden existir diferencias culturales o sociales que lo vuelvan más sencillo para una de estas poblaciones.

Por esto, el análisis y detección de la presencia de DIF en los reactivos de los componentes, es solamente el primer paso en la búsqueda de la interpretación de los resultados de la prueba como imparciales. El siguiente paso implica una revisión de cohorte cualitativo de los reactivos aquí señalados como reactivos con funcionamiento diferencial. Con esta revisión, se podrá verificar si efectivamente existe un sesgo en

la redacción o planteamiento de los reactivos, o si, por el contrario, son indicios de diferencias reales en las poblaciones estudiadas.

Una primera hipótesis que sería importante poner a contraste es la referente a las diferencias en las poblaciones debidas a los contenidos y habilidades requeridas para cada una de las áreas de conocimientos, dadas las similitudes encontradas entre las áreas (CFMI y CBQS) y (CS y HA).

Con todo esto, se concluye que la búsqueda de imparcialidad en los exámenes es una tarea ética fundamental, especialmente al tomar en cuenta el nivel de impacto de las pruebas. Sin embargo, pese a ser una tarea fundamental, es también a su vez una labor inagotable, puesto que cada nueva aplicación requiere de los análisis necesarios para reducir la posibilidad de sesgos.

Recomendaciones

Entre las recomendaciones principales se destaca lo ya mencionado con referencia al análisis cualitativo posterior a la detección de DIF, el cual es recomendable que sea llevado a cabo por expertos en el contenido de los reactivos.

Además de esto, se sugiere dar seguimiento al funcionamiento de los reactivos en futuras aplicaciones. La obtención de una muestra más grande la cual permitirá brindar mayor precisión y consistencia a los análisis.

De igual forma, al contar con aplicaciones de distintos momentos para cada una de las áreas del posgrado, se podrá observar la variabilidad del DIF, lo que a su vez podría permitir identificar qué tan consistente es el funcionamiento diferencial y si pudiera estar dado debido a alguna otra situación externa.

Es importante tomar en cuenta todos los criterios referentes a la calidad y desempeño psicométrico de los reactivos antes de decidir si alguno de estos se utiliza para la calificación o no, e incluso si se es eliminado; pese a la importancia que tiene evitar sesgos a nivel de los reactivos, también es necesario considerar el resto de los indicadores, especialmente tomando en cuenta que existen otras técnicas que permitan subsanar el sesgo en los reactivos una vez que ha sido detectado.

Referencias

AERA. (2014). *Estándares para Pruebas Educativas y Psicológicas*. American Educational Research Association. https://doi.org/10.2307/j.ctvr43hg2

Bond, L., Moss, P., Carr, P. (1996). Fairness in large-scale performance assessment. En: Phillips, G.W., Goldstein, A. (Eds.), *Technical Issues in Large-Scale Performance Assessment*. National Center for Education Statistics, Washington, DC, pp. 117–140.

Cisneros, M., Olave, G. y Rojas, I. (2013). *Alfabetización académica y lectura inferencial*. Ecoe Ediciones.

Dirección de Evaluación Educativa (2017). *Informe del valor predictivo del Examen de Ingreso al Programa Único de Especializaciones en Economía, Generaciones 2015 y 2016. (Convocatorias 2014 y 2015)*.

García-Medina, A.M, Martínez Rizo, F. & Cordero Arroyo, G. (2016). Análisis del funcionamiento diferencial de los ítems del Excale de Matemáticas para tercero de secundaria. Revista mexicana de investigación educativa, 21(71), 1191-1220. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662016000401191&lng=es&tlng=es

Holland, P.y Weiner, H. (1993). *Differential Item Functioning*. Laurence Erlbaun Associates.

Holland, P., Thayer, D. (1988). Differential item performance and the Mantel–Haenszel procedure. En: Wainer, H., Braun, H.I. (Eds.), *Test Validity*. Erlbaum, Hillsdale, NJ, pp. 129–145.

Mair, P. y Hatzinger, R. (2007) Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, 20 (9). pp. 1-20. ISSN 1548-7660

Magis, D., Béland, S., Tuerlinckx, F., y De Boeck, P. (2010). *A general framework and an R package for the detection of dichotomous differential item functioning*. Behavior Research Methods, 42(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Oliveri, M. E., y Wendler, C. (2020). *Higher Education Admissions Practices: An International Perspective*. Cambridge University Press.

Pendfield, R., y Camilli, G. (2006). Differential Item Functioning and Test Bias. En: *Handbook of Statistics: Vol. Psychometrics*. North Holland.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Anexos

Anexo A. Tabla de especificaciones de Comprensión lectora

Tema	Resultado de aprendizaje	Nivel cognoscitivo
Comprensión lectora	Identifica la tesis de un texto.	Comprensión
	Distingue el argumento más importante de un texto.	Comprensión
	Distingue argumentos de apoyo de un texto.	Comprensión
	Comprende el significado de las palabras en el contexto de la lectura.	Comprensión
	Comprende el significado de frases clave en el texto.	Comprensión
	Identifica el tipo de argumentos que se presenta en un texto: de comparación, de causa, ejemplos, deducción y autoridad.	Conocimiento
	Realiza inferencias sobre oraciones o párrafos del texto.	Comprensión
	Identifica las conclusiones del texto.	Comprensión

Anexo B. Tabla de especificaciones de Redacción y gramática

Tema	Resultado de aprendizaje	Nivel cognoscitivo
Reglas gramaticales		
Concordancias gramaticales	Identifica concordancias entre sujeto y verbo.	Comprensión
	Identifica concordancias de género.	Comprensión
	Identifica concordancias de número.	Comprensión
Verbos	Conjuga correctamente los verbos irregulares.	Aplicación
	Usa correctamente correlaciones de tiempos y modos verbales.	Aplicación
	Emplea correctamente los verbos de régimen prepositivo.	Aplicación
	Usa correctamente el gerundio.	Aplicación
Pronombres	Identifica el antecedente de pronombres (personales, demostrativos, posesivos, de objeto directo, de objeto indirecto, reflexivos, recíprocos, "se" impersonal), en oraciones.	Comprensión
Redacción		
Conectores	Emplea los siguientes conectores: preposiciones y frases prepositivas, conjunciones y frases conjuntivas.	Aplicación
Marcadores del discurso	Emplea adecuadamente los marcadores del discurso.	Aplicación
Párrafos	Reconoce párrafos bien construidos (idea completa y estructura adecuada).	Conocimiento
Puntuación	Usa correctamente la coma.	Conocimiento
	Usa correctamente el punto y coma.	Conocimiento
Vocabulario		
Sinónimos	Identifica sinónimos en ejemplos.	Conocimiento
Antónimos	Identifica antónimos en ejemplos.	Conocimiento
Analogías	Establece la relación lógica entre conceptos.	Comprensión
Ortografía		
Uso de s, c, z, sc, b, v, g, j, ll, y, h, r, rr, ü	Usa correctamente las grafías problemáticas.	Conocimiento
Acentuación	Acentúa correctamente.	Conocimiento


UNAM
La Universidad
de la Nación