-

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN PSICOLOGÍA
RESIDENCIA EN EVALUACIÓN EDUCATIVA
FACULTAD DE PSICOLOGÍA

EL PAPEL DEL FUNCIONAMIENTO DIFERENCIAL DEL REACTIVO (DIF) EN LA VALIDEZ DE EXÁMENES DE ALTO IMPACTO

TESIS

QUE PARA OBTAR POR EL GRADO DE:

MAESTRO EN PSICOLOGÍA

PRESENTA:

DAVIN EDUARDO DÍAZ GARCÍA

TUTOR PRINCIPAL:
DR. JOSÉ IGNACIO MARTÍNEZ GUERRERO
FACULTAD DE PSICOLOGÍA, UNAM

MIEMBROS DEL COMITÉ TUTOR:
DRA. MAGDA CAMPILLO LABRENDERO
FACULTAD DE PSICOLOGÍA, UNAM

DR. EDUARDO BACKHOFF ESCUDERO
MÉTRICA EDUCATIVA, A.C.

DRA. CORINA MARGARITA CUEVAS RENAUD
FACULTAD DE PSICOLOGÍA, UNAM

DRA. ALEJANDRA VALENCIA CRUZ
FACULTAD DE PSICOLOGÍA, UNAM

Ciudad Universitaria, CD. MX, Mayo de 2024

Agradecimientos

A mi esposa Caro por motivarme a estudiar este posgrado y por inspirarme a superarme día con día y a no renunciar a mis sueños. Por ser mi fuerza y mi motor y por creer en mí más de lo que yo lo he hecho.

A mi padre por darme ejemplo con su vida de lo que significa entregarse de corazón a una causa, siempre con humilidad, entereza y valentía. Por creer en mí y sentar las bases con su esfuerzo para que pudiera llegar a donde he llegado.

A mi madre, a mis hermanos y a toda mi familia, por forjar en mí los valores humanos y morales que han orientado mi práctica profesional. Y, sobre todo, por siempre creer en mí y por apoyarme en todo momento.

Al Doctor Omar Sánchez-Armass Cappello, por inspirarme y ayudarme a descubrir mi pasión por la psicometría y la programación en R.

Al Doctor José Martínez Guerrero por su valiosa mentoría y apoyo a lo largo de todo el proyecto y por apoyarme a redeescubrir mi pasión por la psicometría.

A la Doctora Lucía Monroy, por ver en mí un potencial que ni yo había visto y por empujarme y motivarme a estar a la altura de ese potencial en cada proyecto.

A todos los colegas psicólogos con quienes he tenido la dicha de colaborar y que a lo largo del camino me han dejado aprendizajes y palabras y acciones de aliento, ayudándome a creer más en mí y mis capacidades.

A CIEES, CONEVAL, MEJOREDU y la DEE de la UNAM por arbirme las puertas dentro de la residencia de la Maestría, ayudándome a consolidar y aterrizar los conocimientos adquiridos dentro del programa.

A la Maestra Nancy y a la Maestra Juliana por confiar en mí y permitirme el acceso a los datos con los que se realizó este trabajo. Sin ellas, esta tesis no habría sido posible.

Índice

Lista de tablas

Lista de figuras

Resumen

La evaluación de sesgos en la medición es una evidencia de validez fundamental en el diseño de exámenes de alto impacto. El funcionamiento diferencial del reactivo (DIF) constituye una de las metodologías más utilizadas para detectar sesgos a nivel de cada reactivo. Dado que existen distintos modelos para detectar DIF, es importante identificar y utilizar los métodos y estrategias que mejor se alinean con los supuestos teóricos y metodológicos que sustenten el examen a evaluar. El presente estudio evalúa la presencia de DIF en dos exámenes, uno de comprensión lectora y otro de redacción y gramática, diseñados como criterio de admisión para programas de posgrado dentro de una de las universidades más importantes de México. Para ello, se utilizaron métodos de detección de DIF paramétricos y no paramétricos, dando especial prioridad a los modelos basados en la Teoría de Respuesta al Reactivo que presentaran un mejor ajuste a los datos según las características de cada examen. Se encontró una mayor presencia de DIF a través de los distintos métodos en tres de los 11 reactivos del examen de comprensión lectora y en cinco de 20 reactivos del examen de redacción y gramática. Los reactivos señalados con presencia de DIF significativo variaron dependiendo del método utilizado, especialmente en el examen de redacción y gramática que resultó mejor representado por un modelo multidimensional. Los resultados se discuten en cuanto a su utilidad teórica y práctica, en la validación de pruebas de alto impacto.

Abstract

The assessment of measurment bias is a foundamental validity evidence when designing high-stake tests. Differential Item Functioning (DIF) is one of the most used methodologies in the detection of item level bias. Since there are different DIF flagging models, it is important to identify and use the methods and strategies that align the most with the theoretical and methodological assumptions that underlie the exam being evaluated. The present study evaluates the presence of DIF in a reading comprehension and a writing and grammar exam designed as admission criteria for graduate programs within one of the most important universities in Mexico. To achieve this, parametric and non-parametric DIF detection methods were used, giving special priority to models based on Item Response Theory that showed a better fit to the data according to the characteristics of each exam. A greater presence of DIF was found through different methods in 3 out of 11 items of the reading comprehension exam and in 5 out of 20 items of the writing and grammar exam. The items identified with significant DIF varied depending on the method used, especially in the writing and grammar exam that was better represented by a multidimensional model. Results are discussed in terms of their theoretical and practical usefulness in high stakes testing.

I. INTRODUCCIÓN

Las Instituciones de Educación Superior (IES) se enfrentan a una serie de retos importantes para cumplir con una enseñanza de calidad; dentro de estos retos, se incluyen cuestiones como la imparcialidad, equidad y oportunidades de aprendizaje (Marta Ferreyra et al., 2017; Silva, 2020). Otro reto importante de las IES se refiere a los procesos de admisión, especialmente cuando hay una competencia por un número limitado de lugares disponibles para el ingreso a la licenciatura o el posgrado (Juarros, 2006). Uno de los enfoques utilizados en estos procesos de admisión es el que se basa en los requisitos y méritos académicos. Dicho enfoque tiene implicaciones importantes con respecto a la validez y a la imparcialidad de los procedimientos de selectividad que se aplican en las universidades y de las decisiones que se toman (Oleas-Falconí y Mosquera-Endara, 2022).

Entre las implicaciones se destaca la validez del uso e interpretación de los resultados de los instrumentos de medición utilizados como criterio de admisión (Sánchez-Mendiola y Delgado-Maldonado, 2017). Una de las interrogantes fundamentales es la capacidad predictiva de dichos instrumentos con respecto al éxito de los alumnos en programas para los que sirven como filtro de admisión (Almarabheh et al., 2022; Makransky et al., 2017). En otras palabras, es crucial presentar evidencia de que los instrumentos utilizados realmente cumplen el objetivo de identificar a aquellos candidatos con mayor probabilidad de culminar exitosamente los respectivos programas de estudios (Lin y Yao, 2014). En ese sentido, y de acuerdo con la Asociación Americana de Investigación Educativa (AERA por sus siglas en inglés), se espera que los instrumentos de evaluación sean, pertinentes en cuanto a sus contenidos, pero también sensibles ante poblaciones de aspirantes, de manera que logren un nivel de imparcialidad que les permita evaluar el nivel de habilidad o atributo que busca medir; y evitar resultados con posibles sesgos provenientede otras variables o aspectos particulares de las distintas poblaciones de sustentantes (AERA et al., 2014).

El concepto de imparcialidad (*fairness*) en el contexto de diseño, elaboración, aplicación y uso de instrumentos en psicología y educación ha sido ampliamente estudiado y ha cobrado mayor relevancia en los últimos años. Este 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 et al., 2014), se reconoce que estudiar la imparcialidad podría implicar abordar desigualdades sociales y otros aspectos no relacionados con la prueba. En términos generales, en los estándares citados, 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 et al., 2014, p. 54)

Para ello, como señaló Osterlind (1983) hace tiempo, es crucial que se garantice a todos los examinados que habrá equidad en los reactivos para asegurar que su habilidad o nivel de desempeño es valorado de forma fiable. De manera tal que, en caso de encontrarse disparidades entre grupos de personas, sean debidas primordialmente a diferencias en la habilidad o desempeño que el test pretenda medir y no a otras variables irrelevantes. Se destaca dentro de las áreas de mayor importancia en este campo la referente al sesgo a nivel de medición de las habilidades que se evalúan. En los Estándares actuales, se reconoce al sesgo de medición como "…una amenaza central a la imparcialidad de la prueba" (AERA et al., 2014). En términos generales, el sesgo de medición se presenta cuando una prueba favorece a un grupo específico por encima de otros. De acuerdo con Bond et al. (1996), podemos entender el sesgo como

> *"… 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 con un estándar de referencia que nos permita conocer el nivel de atributo real para los distintos grupos, y 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 (Penfield y Camilli, 2006). Dada la complejidad de generar este contraste, una alternativa por la que se ha optado es analizar el sesgo de medición a nivel de cada reactivo.

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, se puede decir que existe DIF "… cuando examinandos con iguales capacidades difieren en sus probabilidades de responder a un ítem de la prueba correctamente como una función de pertenencia a un grupo". El análisis DIF tiene gran importancia en este contexto, debido a que es un método que busca identificar posibles sesgos, a fin de evitarlos y asegurar imparcialidad en los resultados de la prueba, para así garantizar que ningún grupo o área específica se encuentre con ventaja o desventaja al responder el instrumento.

Dentro de la postura propuesta por Kane (1992) uno de los principales objetivos es entender la validación como un proceso unificado que incluye distintos tipos de evidencias de validez que presentan un suficiente grado de coherencia. Por ello, para este autor, es importante que los análisis DIF estén insertos dentro de un mismo marco teórico y proceso metodológico que genere un nivel de concordancia suficiente entre las distintas evidencias recabadas sobre un instrumento.

Por ejemplo, si la validez con respecto al correcto funcionamiento de los reactivos y a su dimensionalidad se basó en métodos de la Teoría de Respuesta al Ítem, es importante que los análisis DIF sigan los mismos supuestos y estén enmarcados dentro de la misma teoría y principios.

Por otro lado, si el análisis de la dimensionalidad mostrara que el instrumento es multidimensional, pero el análisis DIF se hiciera bajo el supuesto de que se trata de un instrumento unidimensional, se estarían generando evidencias contradictorias.

No obstante, la importancia de utilizar métodos alineados con el resto de las evidencias de validez, también se reconoce que no en todas las ocasiones es posible optar por métodos robustos como los basados en modelos TRI (Bonifay y Cai, 2017). Hambleton (2006) recomienda contrastar los resultados de distintos métodos para detectar DIF, dado que no todos detectan los mismos reactivos debido a las variaciones en supuestos y procedimientos. Por ello, es importante contrastar distintos métodos para identificar con mayor grado de certeza los reactivos que son señalados con funcionamiento diferencial por diversas metodologías.

II. MARCO TEÓRICO

1. Exámenes de alto impacto

Los exámenes de alto impacto (EAI) son aquellas "[...] pruebas o exámenes cuyos resultados tienen consecuencias importantes y directas para los individuos, programas o instituciones involucradas en el examen" (AERA et al., 2014). Para Menken (2009), en el contexto educativo, un examen se vuelve de alto impacto cuando su solo puntaje se usa como la fuente principal o única para determinar decisiones educativas significativas. Estas consecuencias directas pueden variar ampliamente de efectos positivos a efectos negativos importantes (Sánchez-Mendiola y Delgado-Maldonado, 2017). Stobart (2003), considera que este tipo de exámenes nunca son neutrales, pues siempre tienen consecuencias, ya sea que sean intencionadas o no, o que sean benéficos o perjudiciales.

Dado este impacto directo en las vidas de las personas, resulta imprescindible que los exámenes de alto impacto cumplan con estándares éticos en términos de justicia y equidad (Burgoyne et al., 2021). De manera especial, debido a que, desde su popularización, las críticas a este tipo de evaluaciones han sido constantes, y en muchos casos, bajo el argumento justamente de la inequidad bajo el postulado de que este tipo de evaluaciones puede incrementar la desigualdad en oportunidades (Reeves y Halikias, 2017; Sternberg, 2000).

Algunos de los exámenes de alto impacto más conocidos a nivel mundial incluyen por ejemplo el Examen de Aptitud Escolástica (SAT). El SAT es una prueba de alto impacto administrada por el College Board (2017) en los Estados Unidos. El SAT es un examen estandarizado que mide las habilidades de razonamiento verbal y matemático de los estudiantes que desean ingresar a la educación superior. La prueba consta de secciones de lectura, escritura y lenguaje, y matemáticas. Los resultados del SAT son utilizados por universidades y colegios para tomar decisiones de admisión y otorgamiento de becas. Además, el SAT puede ser una herramienta

valiosa para identificar áreas en las que los estudiantes pueden necesitar apoyo adicional o mejorar sus habilidades (The College Board, 2017).

Otro ejemplo es el Programa para la Evaluación Internacional de Estudiantes (PISA) es un examen de carácter internacional administrado por la Organización para la Cooperación y el Desarrollo Económico (OECD, 2019). PISA evalúa las habilidades de lectura, matemáticas y ciencias de estudiantes de 15 años de edad en más de 70 países y economías. Este examen se realiza cada tres años y proporciona un panorama comparativo del rendimiento educativo a nivel mundial. Si bien podemos clasificar al examen de PISA como uno debajo impacto para los alumnos e incluso para las escuelas, autores como Rivas y Scasso (2021) o Stobart y Eggen (2012) lo catalogan como un examen de alto impacto a nivel gubernamental, dado que los resultados son utilizados por los gobiernos y los responsables de la formulación de políticas educativas para mejorar sus sistemas educativos y abordar las brechas en el rendimiento de los estudiantes (OECD, 2019).

En el contexto mexicano, uno de los ejemplos más importantes es la Evaluación Nacional al Logro Académico en Centros Escolares -ENLACE. De acuerdo con Backhoff y Contreras Roldán (2014), esta prueba fue diseñada inicialmente con propósitos pedagógicos, para más tarde convertirse en una herramienta de rendición de cuentas ante la sociedad. De esta manera,en la práctica se volvió en un examen de alto impacto tanto para escuelas como para maestros; para las escuelas porque se comenzaron a publicar rankings con base en sus resultados, mientras que para los docentes, bonos económicos se introdujeron en función del desempeño de sus alumnos. Estas condiciones, llevaron a que, según los propios Backhoff y Contreras Roldán (2014), el examen ENLACE terminara por corromperse.

Si bien con objetivos específicos diversos, de manera general, las pruebas de alto impacto buscan la mejora continua en los procesos de aprendizaje. Pese a ello, en el contexto real, este tipo de pruebas han generado consecuencias negativas en la vida de los estudiantes, ya sea de forma

directa o indirecta. En una revisión exhaustiva del tema, Minarechová (2012) destaca entre estas consecuencias el estrés, la ansiedad y la discriminación. Además, señala que, la presencia de un "currículo oculto" en las escuelas también influye en cómo los estudiantes enfrentan y experimentan el éxito o fracaso en estas pruebas, y esto puede variar según el género. A su vez, en los exámenes con alto impacto para las escuelas, algunos estudiantes pueden ser retenidos, expulsados o reubicados para mejorar artificialmente los resultados de la escuela (Minarechová, 2012).

Además de los impactos negativos en las vidas de los estudiantes, para Minarechová (2012), algunos exámenes de alto impacto para las escuelas también pueden tener un impacto negativo en la enseñanza y el aprendizaje. Uno de los principales problemas es que, al enfocarse en el mero objetivo de mejorar los puntajes para la escuela, los maestros tienden a centrarse en estrategias de enseñanza basdas en la memorización o en la práctica de resolución de reactivos, en lugar de enfoques más innovadores, como el aprendizaje cooperativo y los proyectos creativos (Blazer, 2011). Además, las pruebas de alto impacto pueden llevar a un mayor énfasis en el contenido evaluado, en detrimento de otras áreas del currículo. Esto puede resultar en la marginación de asignaturas que no están incluidas en las pruebas, como arte, educación física, estudios sociales y ciencias (Amrein y Berliner, 2002; Westchester Institute for Human Services Research, 2003).

El efecto negativo de las pruebas de alto impacto también se extiende a los profesores y al proceso de evaluación. Por ejemplo, muchos maestros han abandonado la profesión porque no se encontraban de acuerdo con las políticas implementadas a partir de la gran importancia dada a los exámenes de alto impacto, ya sea por una gran presión por mejorar resultados de sus alumnos o por la implementación de métodos de enseñanza rígidos enfocados en los contenidos de dichos exámenes (Amrein y Berliner, 2002). Estos problemas, junto con los costos significativos asociados con la implementación de pruebas, plantean preguntas sobre si los fondos se están utilizando de la

mejor manera posible y si estas pruebas realmente benefician a los estudiantes y a la educación (Baines y Stanley, 2005).

2. Procedimientos de ingreso al posgrado

En México, la demanda por programas de posgrado ha ido en aumento a lo largo de las últimas décadas (Álvarez Montero et al., 2014), una de las bases y pautas más importantes y frecuentes es la meritocracia (Sternberg, 2010). Bajo esta premisa, el acceso a la educación superior, especialmente en situación con alta demanda y poca oferta de lugares, se otorga dando preferencia a aquellos mejor preparados para transitar por esta fase de estudios. Para llegar a esta determinación en términos de distinguir y elegir a aquellos con la suficiente capacidad, cada universidad decide utilizar sus propias estrategias y procedimientos de admisión, los cuáles suelen depender mucho del contexto propio de la institución (Oliveri y Wendler, 2020). Entre los métodos más utilizados se encuentran el tomar como referencia calificaciones del último grado obtenido, entrevistas, revisión de currículos y exámenes de conocimientos generales como filtro para el ingreso a una institución o programa educativo (Sternberg, 2010).

Si bien existen alternativas recientes entre estas metodologías, como la consideración de diversidad a nivel étnico y de nacionalidad (Bastedo, 2021), el uso del resto de técnicas sigue siendo la norma. Por ello, se supone que estos métodos permiten determinar qué estudiantes tienen mayores probabilidades de concluir sus estudios. Por lo cual, es importante contar con evidencia respecto a la capacidad predictiva medida en todos estos métodos.

Dentro de las áreas más utilizadas en exámenes de alto impacto para la admisión a programas de estudios universitarios, la medición de habilidades verbales ha cobrado particular importancia, siendo esta una de las áreas más incluidas en pruebas de alto impacto como el SAT ya mencionado con anterioridad (The College Board, 2017).

3. Medición de habilidades verbales

En el caso de los exámenes de alto impacto que se enfocan en las habilidades verbales, McNamara et al. (2019) mencionan que suelen ser exámenes sumamente ambiciosos pues con frecuencia presuponen que recopilan evidencia para generar un panorama lo suficientemente completo sobre las habilidades verbales de la persona sustentante.

Sin embargo, para abordar el tema de habilidades verbales, es necesario comenzar por hablar del lenguaje; tal como lo señala el Grupo de Trabajo Conjunto sobre Evaluación de la Asociación Internacional de Lectura y el Consejo Nacional de Profesores de Inglés (IRA y NCTE, 2009), el lenguaje es como un organismo vivo que existe solo en la interacción con otros, sin tener un significado por sí mismo, sino uno que se construye en la relación social donde se le usa. De manera interesante, este grupo de trabajo conjunto señala que las personas dan sentido al lenguaje en este contexto relacional, tomando en cuenta tanto su historia personal como su memoria colectiva, utilizando esta información para dar sentido a una sola palabra o a una frase.

Precisamente por esta complejidad y subjetividad contextual del lenguaje, es importante ser cautelosos con el uso e interpretaciones previstas en exámenes que evalúan habilidades verbales, dado que, si bien el lenguaje es el objeto de evaluación, también es el medio por el que se evalúa. Es en ese mismo sentido que la IRA y NCTE (2009) señalan que “[...] cuando intentamos estandarizar un examen (hacerlo igual para todos), hacemos la tenaz suposición de que todos los estudiantes darán el mismo significado al lenguaje que utilizamos en las instrucciones y al lenguaje de cada uno de los reactivos". Esta suposición da lugar a sesgos que resultan inevitables, y que por lo mismo, resulta imperativo establecer control sobre éstos en el uso de exámenes, especialmente en aquellos de opción múltiple (Lions et al., 2021).

Las habilidades verbales son un constructo complejo de definir, especialmente en términos de los componentes o dimensiones que lo constituyen. Según Thirakunkovit (2018), el concepto de

habilidades verbales puede interpretarse tanto de manera unidimensional como multidimensional. Mislevy y Yin (2013) sugieren que al hablar de habilidades verbales en evaluaciones, se refleja la intención de interpretar aspectos específicos de las capacidades y competencias de los examinados.

En ese sentido, la naturaleza del constructo que se pretende evaluar varía de manera significativa dependiendo del caso y la aproximación utilizada (Baehman, 2007), destacando principalmente tres aproximaciones teóricas diferentes: la perspectiva de rasgo, la conductista y la interaccionista. De acuerdo con Chapelle (1998) estas perspectivas se pueden definir en términos de cómo explican la consistencia en las respuestas de los sustentantes.

Según la perspectiva de rasgo estas consistencias se atribuyen a características específicas de los sustentantes, siendo sus respuestas manifestaciones de rasgos latentes. Por otro lado, desde la perspectiva conductista se atribuye la consistencia de las respuestas completamente a factores contextuales, por lo que hacen especial énfasis en las condiciones ambientales bajo las que se dan las respuestas. Finalmente, la perspectiva interaccionista entiende la consistencia de respuestas tanto a partir de rasgos del sustentante como de factores contextuales y su interacción (Chapelle, 1998)

Esta distinción de perspectivas tiene un impacto considerable en la construcción de evaluaciones, especialmente en términos de las interpretaciones que se harán de los resultados. Desde una perspectiva de rasgo, dado que las respuestas son reflejo de una estructura latente, se asume que el nivel de habilidad medido en la prueba es fácilmente trasladable a otros contextos, mientras que bajo la perspectiva conductual, se da un especial énfasis en verificar la semejanza entre los reactivos y las situaciones del mundo real, para poder ver el alcance de su generalizabilidad (Mislevy & Yin, 2013).

La distinción entre estos posicionamientos epistemológicos es trascendental en el proceso de desarrollo de exámenes de habilidades verbales, puesto que marca la pauta de su construcción y

conceptualización general. Por ello, es importante destacar que el consenso general de psicómetras y desarrolladores de pruebas se ha inclinado más hacia una visión interaccionista. Por ejemplo, Chapelle (2013) señala que ya no se toma en serio la idea de que las habilidades verbales pueden encasillarse en un solo constructo unidimensional, y que la cuestión es más bien la utilidad de los modelos que se utilizan para representarlas. Desde esta perspectiva pragmática, dado que los modelos siempre son una simplificación de los fenómenos, el foco de atención no está en qué tan precisos son al representar la realidad, sino en qué tan útiles resultan para los propósitos de los exámenes.

En última instancia, como señalan Mislevy y Yin (2013), "Un constructo en una evaluación es la concepción inevitablemente simplificada del evaluador de las capacidades del examinado, elegida para adecuarse a un propósito de una evaluación dada" (p. 215). Por esto, es importante verificar el modelo que mejor ajusta a los datos en términos de dimensionalidad empírica y teórica.

3.1 Evaluación de la comprensión lectora

En el campo específico de la lectura o comprensión lectora, Weir (2005) señala que existe un debate continuo sobre la naturaleza del constructo, entre quienes argumentan que se trata de un proceso unitario e indivisible o quienes lo dividen en componentes de habilidades u operaciones distinguibles entre sí. Quienes ven a la comprensión lectora como un constructo unidimensional argumentan que evaluar los componentes de manera aislada no nos otorga una medida verdadera del constructo en sí mismo (Weir, 2005).

Por su parte, Grabe y Stoller (2019) señalan la dificultad de encasillar la comprensión lectora en una definición universal del estilo "habilidad para extraer significado de la página impresa e interpretar esta información apropiadamente". Principalmente debido a que esta definición no contempla aspectos como las diferentes formas de aproximarse al texto dependiendo del propósito que tenga el lector para hacerlo y que no se resaltan todas las habilidades, procesos y

conocimientos previos necesarios para llevar a cabo la lectura ni considera el contexto social en que se realiza la lectura ni los usos que se le dará a la interpretación.

No obstante la complejidad de este debate, Weir (2005) señala que, independientemente de la posición teórica que tomemos, el énfasis debería de estar en el respaldo de las aseveraciones que hagamos sobre qué estrategia o habilidades específicas mide qué reactivo en un examen. Señala finalmente que: "[...] si queremos reportar la competencia del estudiante en comprensión lectora como un constructo distinto de, por ejemplo, habilidad de redacción, nos vemos forzados a diseccionar la comprensión lectora en las que hipotetizamos con sus componentes constituyentes" (p. 88), de tal manera que podamos evaluar dichos componentes y utilizar los resultados compuestos para reportar la capacidad lectora.

Propuestas sobre los componentes que se suponen constituyentes a la lectura existen muchas y muy diversas, y aunque una buena parte de ellas se concentran en la lectura de un segundo idioma (e.g., Urquhart y Weir, 1998) siguen incluyendo conceptos que son generalizables a todo proceso de lectura. En este trabajo, nos concentramos en las propuestas de Weir et al.(2000) y de Grabe y Stoller (2013).

Weir et al. (2000) subrayan la distinción fundamental entre estrategias y habilidades. Definen estrategias como acciones conscientes dirigidas a resolver problemas, mientras que las habilidades son procesos menos conscientes, ejecutados en su mayoría de manera automática. En contraste, Grabe y Stoller (2013) señalan que esta separación puede volverse menos clara, ya que algunas estrategias pueden automatizarse. En su perspectiva, las estrategias son habilidades con potencial para ser controladas conscientemente, enfatizando que comprender esta distinción mejora al considerar el propósito específico de la lectura.

Weir et al. (2000) no hablan de manera central sobre el propósito u objetivo por el que se lee, más bien se enfocan en hacer la distinción entre la lectura a nivel global o local, donde en el

primer caso, se comprende a nivel macro, destacando aspectos como las ideas principales; mientras que, en el segundo caso, se refiere a la comprensión a nivel micro, haciendo énfasis en el significado léxico de palabras específicas, referencias pronominales, entre otras. Sin embargo, Grabe y Stoller (2013) sí enfatizan el propósito de leer como un proceso central para conceptualizar la lectura. Para ellos, existen siete propósitos principales por los que una persona podría aproximarse a la lectura y cada uno de ellos implica estrategias o habilidades diversas:

- *Buscar información simple:* Para este fin, normalmente escaneamos el documento buscando una palabra específica o información particular
- *Hojear rápidamente:* Involucra una serie de estrategias específicas que llevan a suponer en qué partes del texto hay información importante, para obtener una idea general tras leer estas secciones identificadas.
- *Aprender del texto:* Requiere de habilidades como el recordar ideas principales y otros detalles que den soporte a éstas y a ideas secundarias, reconocer y estructurar marcos retóricos que permitan organizar la información y relacionar el texto con el conocimiento previo del lector.
- *Integrar información, buscar información necesaria para escribir o criticar el texto:* Requiere habilidades adicionales a las del punto anterior, tales como la evaluación crítica de la información leída, permitiendo determinar qué información y de qué manera la integra el lector para su propósito específico.
- Comprensión general: Pese a ser el propósito más habitual, es uno de los más complejos. Requiere un procesamiento rápido y automático de la información, habilidad para formar una representación general del significado de las ideas principales y una coordinación eficiente de distintos procesos dentro de un marco limitado de tiempo.

Para Grabe y Stoller (2013), estos propósitos pueden tener múltiples variantes y combinaciones, por lo que no son una regla general o un sistema de clasificación absoluto. En última instancia, como destaca Weir (2005), se trata de identificar los procesos que consideramos importantes incorporar en un examen y que sea cercano a la conceptualización que se hace sobre lo que significa la comprensión lectora; aunque esto no implique que el examen sea un modelo 100% válido de lo que es la comprensión lectora. Por ello, es importante que los reactivos y los formatos utilizados en un examen estén alineados con el propósito.

En relación con los formatos de respuestas para evaluar comprensión lectora, Weir et al. (2000) recomiendan utilizar distintos métodos para la medición de este constructo, exhortando incluso al uso de formas de evaluación más subjetivas a modo de complemento de las técnicas objetivas tradicionales. Pese a ello, (Weir, 2003) señala que es muy importante cuidar que las técnicas utilizadas no terminen por interferir con la medición del constructo. Como se señaló en secciones anteriores, al ser el lenguaje el objeto evaluado y a la vez el medio para su evaluación es posible que las técnicas utilizadas generen ruido innecesario.

Por ejemplo, puede darse el caso que, si bien el estudiante comprenda el texto, no termine por entender lo que el reactivo está demandando, en caso de que su redacción no sea adecuada. Además de la claridad en la redacción de los reactivos, para Weir (2003) resulta fundamental cuidar también los formatos de respuesta. En ese sentido, reactivos de opción múltiple podrían ser una opción favorable, pues no entorpecerían la medición de la comprensión lectora solicitando al sustentante que escriba su respuesta.

Los exámenes de opción múltiple tienen la ventaja de permitir una focalización más específica de lo que se espera del examinado. Es posible seleccionar el texto o una parte de él para cada pregunta, lo que delimita los procesos y habilidades requeridos para responder. Esta aproximación contribuye a establecer una validez respaldada por la teoría (Weir, 2003).

3.2 Evaluación de la escritura

El constructo referente a la habilidad de escritura incluye algunos matices particulares que lo distinguen dentro de las habilidades verbales. De acuerdo con Hyland (2021), la escritura puede ser abordada desde aproximaciones muy distintas, desde aquellas que se centran en el texto como producto hasta las que lo hacen en el escritor y su proceso de creación e incluso aquellas que enfatizan el papel del lector y el proceso de interpretación del texto. En el primer caso, que sería el foco de esta investigación, el autor señala también distintos aspectos que se pueden enfatizar dentro de esta aproximación. Por ejemplo, es posible ver al texto como un objeto analizable en términos de retórica, de recursos lingüísticos, estructura e incluso de reglas gramaticales. Por otro lado, es posible ver al texto más como un discurso, enfocado más bien en el propósito comunicativo y la manera en que el lenguaje es utilizado para alcanzar ese fin.

En ambos casos, el análisis y evaluación de la habilidad de escritura partiría necesariamente de la evaluación de un producto directamente, donde la principal técnica de medición es la evaluación de un portafolio. Si bien es cierto que este tipo de evaluación presenta bondades importantes, puesto que permite una evaluación más cercana al contexto real, facilitando hacer interpretaciones más generalizables, Weir (2003) destaca que existen serias dudas respecto a los procesos de calificación de portafolios, que pueden tener un impacto en la validez de las interpretaciones realizadas.

Para Weir (2003), la validez basada en evidencia al evaluar escritura se enfoca en medir la activación de recursos y procesos ejecutivos que son evocados por la tarea; dentro de estos recursos destaca los recursos lingüísticos y los de conocimiento del área sobre la que se escribe. En ese sentido, este autor señala que se han hecho intentos para dividir la escritura en elementos microlingüísticos discretos y más específicos, como la gramática, ortografía, vocabulario o puntuación. El autor señala que este tipo de exámenes serían una medición indirecta del constructo

de escritura, pues solamente medirían partes o recursos que consideramos constituyen el proceso de escritura, sin acercarse a medir el constructo completo de forma más directa.

Uno de los aspectos más simples incluidos dentro de estos elementos microlinguísticos discretos es el de la ortografía, aunque, de evaluarse de forma aislada, sería una reducción de la escritura a aspectos meramente gráficos. Como bien señala Morales Ardaya (2004), de nada sirve un texto impecable a nivel de ortografía si se compone de oraciones confusas sin concordancia gramatical y que impiden su comprensión. Aun así, este mismo autor señala que, no por ello la ortografía resulta prescindible en cualquier evaluación de la escritura, ya que sigue siendo una parte fundamental de ella, pero que necesariamente debe de ser acompañada de otros componentes más complejos. Entre estos, Morales Ardaya (2004) destaca en particular el vocabulario, la gramática, la cohesión y la estructura del texto. Este último punto resulta complejo de evaluar en exámenes de opción múltiple, por lo que el foco estará en los demás componentes.

El vocabulario es otro de los subcomponentes del proceso de escritura que ha sido ampliamente estudiado. Algunos autores se refieren a este como competencia léxica, ya que implica más que el simple conocimiento memorístico de la palabra. De acuerdo con (Jiménez Catalán, 2002), la competencia léxica implica tanto el conocimiento necesario para poder usar la palabra de forma correcta como la capacidad de reconocer, aprender, recuperar y relacionar las distintas palabras a nivel oral y escrito.

El conocimiento gramatical entendido como “las correspondencias entre forma y significado” (Purpura, 2012), es otro componente esencial en el proceso de redacción y en general de la evaluación de habilidades verbales. En el contexto de habilidades verbales en un segundo idioma, Purpura (2012) destaca que el conocimiento gramatical surge invariablemente en cualquier uso del lenguaje y que la evaluación de este componente nos brinda información sobre las formas

gramaticales que los estudiantes son capaces de utilizar al intentar expresar significados en un contexto específico.

Aún considerando las ventajas que esta aproximación puede tener en términos de validez y del proceso de calificación, es importante ser cautelosos en su uso, puesto que cuentan también con algunas limitaciones importantes, entre las que Weir (2003) destaca las siguientes:

- Pese a estar relacionados con la capacidad de escritura, no pueden representar en su totalidad todo lo que un escritor con alto nivel de habilidad puede hacer
- Involucran pocos de los recursos y procesos que son específicos de la escritura como constructo general, por lo que pueden ser limitados en términos de su validez basada en la teoría
- Resulta más complejo generalizar las interpretaciones y pretender estimar a partir de los puntajes qué tan bien los sustentantes se desempeñarán en contextos más reales.

A esto hay que sumarle lo que Hyland (2009) señala en tanto que, si bien estas pruebas pueden ser una fuente confiable y precisa en su medición, no toman en cuenta que el objetivo de la escritura es la comunicación, y no la precisión en términos de gramática u ortografía. Por todo esto, es importante delimitar las interpretaciones que se harán de este tipo de exámenes, y cuidar el alcance de posibles generalizaciones que son posibles dada la estructura y forma de la evaluación.

4. Validez

Dada la complejidad de los constructos de habilidades verbales y las múltiples definiciones e interpretaciones que de ellos surgen, se vuelve crucial, desde un posicionamiento metodológico, ético y profesional dentro del proceso de elaboración de exámenes que buscan valorar estas habilidades, considerar como foco de atención el concepto de validez. Para Alderson et al. (1995), la ética en la evaluación de habilidades verbales no es más que una validez extendida. Y es que, como

señala Davis (2008), no es factible considerar todas las posibles consecuencias sociales de un examen, solo podemos ocuparnos de consecuencias limitadas y predecibles, haciéndonos responsables de ellas. Este proceso de delimitación y anticipación de consecuencias es entendido por Messick (1989) como la validez de consecuencias. Esta visión de la validez es más cercana a la visión contemporánea, y es importante generar esta distinción para comprender el proceso de una forma más integral.

De acuerdo con la revisión realizada por Chapelle (2013), el concepto de validez ha estado presente desde la década de 1920, y muchas de las menciones hasta el año de 1971 lo definen en términos de una propiedad del examen, refiriendo que es válido si realmente mide lo que pretende medir. Esta visión ha permeado y perdurado por varias decadas, ya que muchos textos de carácter didáctico siguieron utilizando este tipo de definiciones (Davies, 1990; Hatch y Lazaraton A., 1997). Incluso, en el lenguaje utilizado, pues existen muchas referencias a procesos de validación de instrumentos, o menciones de que cierto instrumento "está validado". Para Chapelle (2013), estas menciones refuerzan la visión de que la validez es una propiedad del instrumento y liberan a los usuarios de su responsabilidad sobre la validez de la interpretación y uso de los resultados.

De acuerdo con Zumbo (1999), la visión clásica de la validez considera si un instrumento está midiendo lo que pensamos que mide y resulta ser una propiedad misma del instrumento, normalmente definida a partir de estadísticos muy puntuales, tal como la correlación con otras pruebas. Desde esta perspectiva, la validez se vuelve determinística en el sentido de que se asevera simplemente si un instrumento es o no válido. Cabe mencionar que si bien, de esta manera la validez se convierte en algo categórico, desde esta visión clásica también se le caracteriza en diferentes grados y distitno tipos tales como la validez de contenido, de criterio o de constructo.

Pese a la persistencia de este tipo de conceptualizaciones tradicionales referentes a la validez, el consenso general dentro de la psicometría se ha movido hacia una visión más compleja

sobre este concepto, siendo que, por ejemplo, la AERA define la validez como "[...] el grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas." (AERA et al., 2014).

A partir de esta definición, podemos destacar algunas características importantes de esta visión de la validez que difieren significativamente de la perspectiva clásica. Dado que la validez es referente a las evidencias que respaldan las interpretaciones de los puntajes de una prueba; Weir (2005) señala que podemos decir que la validez se refiere a los puntajes en una administración específica y no a la prueba en sí misma.

Por otro lado, dado que el foco está en el grado en que la evidencia sirve de respaldo a las interpretaciones de los puntajes, Messick (1993) enfatiza que la validez es justamente un aspecto gradual, más que en términos absolutos de presencia o ausencia, siendo así más bien un concepto relativo. Además de esto, dado que se requiere evidencias que respalden las interpretaciones y usos previstos, la validez no reside en lo que los autores del test aseguren, sino en las evidencias proporcionadas que respalden esas aseveraciones.

Finalmente, dado que la validez incluye aspectos relacionados con los usos específicos que se pretende dar a los puntajes de la aplicación del instrumento, necesariamente esto incluye un componente de consecuencias sociales (Hubley y Zumbo, 2011), haciendo así el concepto mucho más amplio que la visión clásica donde se le veía como un concepto meramente técnico. En realidad, la complejidad de la visión actual hace a la validez un proceso continuo que no es agotable, pues todo uso e interpretación previsto requiere y puede tener distintos tipos de evidencia y distintos grados de respaldo a partir de la solidez de ésta.

En ese sentido, una de las propuestas más completas y que permiten una visualización más integral del concepto de validez es el enfoque argumentativo. Este enfoque, en palabras de su autor, es "[...] un marco sustentado en la interpretación para recolectar y presentar evidencia válida y

explícita asociada con la plausibilidad de varias suposiciones e inferencias involucradas en la interpretación" (Kane, 1992, p. 2). A partir de esto, este enfoque vuelve necesarios dos requisitos fundamentales: declarar abiertamente las aseveraciones que se harán sobre el test en términos de usos e interpretaciones, y evaluar la credibilidad de dichas aseveraciones (Kane, 2021)

Al explicitar las interpretaciones y usos propuestos, se explicita también las inferencias que se están realizando, las cuales adoptan una forma de "si-entonces". Por ejemplo, "si el sustentante obtiene un puntaje mayor a un punto de corte establecido, entonces es interpretado como competente". Para Kane (2021) solo es posible considerar como razonables estas inferencias si son altamente probables a priori o si se sustentan en la suficiente evidencia.

La primera opción es importante, pues para este autor, toda inferencia realizada suele sustentarse en una gran cantidad de suposiciones, de las cuales, muchas suelen simplemente ser asumidas como verdades sin requerir mayor tipo de evidencia. El ejemplo que da Kane (2021) es que, al hacer un examen de admisión, nosotros suponemos que los sustentantes comprenden las instrucciones y las preguntas, y que son capaces de registrar sus respuestas. Bajo este enfoque, solo cuando existan dudas significativas sobre las suposiciones realizadas es que resultará importante aportar evidencia que las sustente.

Por ello, diferentes inferencias requieren de distintos tipos de evidencia, al punto que aseveraciones o inferencias con un mayor impacto en las vidas de los sustentantes, más sólidas y robustas deben de ser las evidencias presentadas para respaldar esa aseveración o inferencia, al punto que las suposiciones más cuestionables son las que mayor atención deben de tener en el proceso de presentación de evidencias de validez (Kane, 2021). Para lograr este cometido, es importante tomar una postura crítica con respecto a estas aseveraciones. De acuerdo con Cronbach (2013), el proceso de validación no se trata de respaldar interpretaciones, si no de encontrar las flaquezas que puedan tener.

En ese sentido, para Messick (1989), lo que se ha de cuestionar debe ir más allá de las evidencias empíricas y teóricas sobre las inferencias realizadas, se debe analizar incluso, los valores de los investigadores que se encuentran detrás de sus interpretaciones y usos previstos para las pruebas. Para él, "dada la omnipresencia y sutileza con la que los valores e ideologías impactan las interpretaciones de las pruebas, debemos explorar formas y medios para destapar premisas de valor tácito y de lidiar con sus consecuencias para la validación de pruebas" (p. 104).

Chapelle (2013) presenta algunos ejemplos importantes de la influencia que los valores del investigador y de los interesados en los exámenes pueden tener en los usos e interpretaciones previstos. Destaca por ejemplo cómo se usa a los hablantes nativos de un idioma como la norma y la referencia general, o como puede tratarse igual a personas que aprendieron un idioma por herencia, desde la escucha en casa y a aquellos que lo aprendieron dentro del aula. Para esta investigadora, esto es un reflejo de cómo desde la operacionalización y definición de constructos se puede estar generando una ventaja importante para ciertos grupos. Finalmente señala sobre este punto que, dado que los constructos de habilidades verbales pueden definirse de forma pragmática y desde muchas aproximaciones, es importante analizar los valores de fondo para ver las implicaciones que puede tener la definición un constructo desde distintas bases y enfoques.

5. Imparcialidad

A partir de esta conceptualización, y luego de haber señalado los peligros de definir constructos desde una postura de valor posiblemente inclinada a beneficiar a grupos específicos, es trascendental discutir el tema de la imparcialidad dentro de los exámenes de habilidades verbales, pues tomada como un valor de inicio desde un posicionamiento ético y profesional, puede ayudar a generar exámenes más justos.

La imparcialidad ha ido cobrando mayor relevancia a lo largo del tiempo; tal como señala Worrell (2016), dentro de los propios Estándares de la AERA, el concepto ha ido ganando terreno,

de ser simplemente mencionado en las primeras ediciones a tener un capítulo completo dedicado a ella en la más reciente edición, convirtiéndolo así, desde la perspectiva de Worrell (2016) en el *tercer grande*, uniéndose en relevancia e importancia la validez y la confiabilidad. De manera puntual, los Estándares definen 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, p. 54).

Esta definición es después desglosada dentro de los Estándares en distintas aproximaciones que cubren dichas características individuales y contextuales de la evaluación; dentro de ellas, las dos de mayor importancia para el presente trabajo son las referentes al sesgo de medición y la referente a las evidencias de validez de imparcialidad para los usos previstos. En este sentido, tal como lo argumentan McNamara et al. (2019), la imparcialidad es una extensión y parte del proceso de evidencias de validez, volviéndose un componente específico de la prueba y su proceso de aplicación. Estos autores hacen la distinción entre imparcialidad y justicia.

Para ellos, la justicia hace referencia a los usos de los exámenes que se determinan de forma externa, refiriéndose de forma específica al componente social y político que involucra todo proceso de diseño, evaluación y aplicación de exámenes. Dentro de este trabajo, el foco estará más bien en el concepto de imparcialidad tal como lo definen estos autores, como un componente adicional del proceso de validez (McNamara et al., 2019).

Esta postura es concurrente con la de los propios estándares, pues para la AERA et al. (2014), la imparcialidad se trata de identificar y posteriormente eliminar lo que ellos llaman barreras que imposibilitan el desempeño óptimo de los sustentantes y que son irrelevantes al constructo. Al eliminar dichas barreras, se estaría permitiendo una interpretación más válida y comparable de los puntajes de todos los sustentantes.

Se destaca dentro de las áreas de mayor importancia la referente al sesgo a nivel de medición, reconocido dentro de los estándares como "[…] una amenaza central a la imparcialidad de la prueba" (AERA et al., 2014, p. 54).

6. Sesgo de medición

En términos generales, el sesgo de medición se presenta cuando una prueba favorece a grupos específicos por encima de otros. De acuerdo con Bond et al. (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).

Para McNamara y Roever (2006), esto significa que la pertenencia a un grupo específico introduce varianza sistemática que es irrelevante al constructo. Esta distinción es importante, pues como los mismos autores señalan, si bien cualquier varianza irrelevante al constructo puede ser problemática, cuando se presenta de forma sistemática termina por perjudicar de forma constante a un mismo grupo, infringiendo así los principios de imparcialidad.

El sesgo es de tal importancia para la validez de una prueba que, en su momento, (Shepard et al., 1985), p. 179) definió al sesgo de medición simplemente como "invalidez", dada la fuerte relación que existe entre imparcialidad y validez, donde cuando una incrementa, necesariamente la otra decrementa.

Dentro de este sesgo de medición, Van de Vijver y Poortinga (2004), en el contexto de evaluaciones transculturales, señalan que existen tres tipos de sesgos: sesgo de constructo, metodológico y sesgo de ítem.

El sesgo de constructo ocurre cuando el constructo medido no es idéntico en diferentes grupos culturales. Esto puede suceder si una prueba mide diferentes rasgos o habilidades en distintos grupos culturales o cuando el rasgo o habilidad medido no es igualmente relevante o importante en todos los grupos culturales.

El sesgo de método se refiere a las diferencias en la forma en que se obtienen las puntuaciones de las pruebas que no están relacionadas con el rasgo o habilidad medido. Esto puede incluir tanto diferencias en la administración de la prueba, como diferencias en las instrucciones o el entorno físico, así como diferencias en los estilos de respuesta o estrategias para tomar la prueba.

El sesgo de ítem ocurre cuando los ítems individuales en una prueba funcionan de manera diferente para diferentes grupos culturales. Esto puede suceder cuando un ítem es más difícil o más fácil para un grupo en comparación con otro, incluso después de controlar las diferencias en los niveles de habilidad.

7. Funcionamiento Diferencial de los Reactivos (DIF)

De acuerdo con Penfield (2016, p. 56): *"Dado que el puntaje final del test se basa en la acumulación de evidencias de cada elemento que se puntúa en las respuestas de los examinados, se puede desarrollar un argumento de imparcialidad en los puntajes de las pruebas a partir de la evaluación de factores irrelevantes al constructo asociados con los elementos individuales que se puntúan para generar el puntaje de la prueba."* De acuerdo con este mismo autor, esta evidencia de que cada elemento está libre de factores irrelevantes al constructo resulta convincente como prueba de que el puntaje global lo está también.

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 sustentantes 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. El análisis DIF

tiene gran importancia en este contexto, ya que es un método que busca identificar posibles sesgos y con ello fomentar imparcialidad en la aplicación de la prueba para garantizar que ningún grupo o área específica parta con ventaja al responder el instrumento. La presencia de DIF puede impactar de gran forma la validez de la prueba, la imparcialidad y la posibilidad de comparación de puntajes entre los distintos grupos evaluados (Cheng et al., 2020). Existen muchas y muy diversas metodologías para detectar la presencia de DIF, algunas de ellas directamente derivadas de la Teoría de Respuesta al Ítem (TRI), y algunas otras de los principios y supuestos de la Teoría Clásica de los Tests (TCT).

7.1 Métodos para la detección de DIF

Un marco general de las posibles categorizaciones de métodos para estimar DIF es el ofrecido por Magis et al. (2010). De acuerdo con estos autores, son cuatro las dimensiones más importantes que debemos considerar para identificar y clasificar las metodologías existentes para el análisis DIF:

1. Número de grupos focales: Dentro de la terminología utilizada en la literatura referente a DIF, se suele distinguir entre grupos focales y grupo referencial. Los grupos focales son aquellos que se espera se encuentren en desventaja con respecto a los posibles sesgos dentro del examen, mientras que el grupo de referencia suele ser aquel para el que el examen se planteó inicialmente o para el que se espera funcione de manera adecuada.

Un ejemplo de ello son los exámenes de inglés, donde el grupo de referencia suele ser las personas cuya lengua materna es ésta, y dentro de los grupos focales se suele incluir a personas con una lengua materna distinta al inglés. El caso más común es tener un grupo de referencia que se desea comparar con un grupo focal. Sin embargo, puede que se requiera evaluar a más de un grupo focal, o, inclusive, que no se tenga un grupo de referencia específico pero que aun así se desee comparar dos o más directamente. La mayoría de los métodos clásicos en el análisis DIF se

diseñaron para comparar dos grupos, aunque algunos de estos métodos se han generalizado para permitir el análisis de múltiples grupos.

2. Aproximación metodológica: De acuerdo con Magis et al. (2010), podemos clasificar los métodos existentes a partir de si utilizan un modelo TRI como base o no. En el primer caso, para realizar un análisis DIF, se necesita estimar un modelo TRI como primer paso. En el segundo caso, el análisis DIF se basa en métodos estadísticos para datos categóricos, usando como criterio principal el puntaje total de la prueba.

3. Tipo de efecto DIF: El efecto DIF se refiere a las diferencias en las probabilidades de responder el reactivo de manera correcta. Si dichas diferencias en la probabilidad son independientes del nivel de habilidad de los sustentantes, se dice que el efecto DIF es uniforme. Si, por el contrario, las diferencias en probabilidad no son constantes a través del nivel de habilidad (e.g., las personas con mayor nivel de habilidad de un grupo se ven beneficiadas con el reactivo, pero las de menor nivel de habilidad no), se dice que el efecto DIF es no uniforme o cruzado.

4. Purificación de reactivos: Este punto, más que un medio de clasificación se refiere a un paso adicional que se puede utilizar para la obtención del efecto DIF. Dado que el marco de referencia utilizado para contrastar la presencia de DIF es el puntaje total de la prueba en los modelos no paramétricos, y el nivel de habilidad (calculado con el conjunto total de reactivos) en los modelos paramétricos, puede ocurrir que la presencia de DIF en un reactivo influya en la detección de DIF en otros reactivos., pues el punto de referencia estaría en sí mismo sesgado.

Por ello, este procedimiento implica detectar el efecto DIF en el conjunto total de reactivos, para posteriormente comenzar a analizarlos uno por uno sin tomar en cuenta los demás reactivos que presentaron DIF. Esto ayuda a evitar la influencia del DIF presente en algún reactivo, en su detección en otros.

A continuación, se presenta en la Tabla 1, un resumen de los métodos más comunes utilizados para cada uno de los casos descritos en el punto anterior que fue recuperado de Magi et al. (2010). Cabe aclarar que los modelos aquí presentados son para reactivos de tipo dicotómico exclusivamente.

Tabla 1.

Propuesta de clasificación de métodos para detectar DIF

Marco teórico	Tipo de efecto DIF	Número de grupos	
		2	>2
No paramétricos	Uniforme	Mantel–Haenszel SIBTEST Regresión Logística	Mantel–Haenszel generalizado Comparación por pares mediante SIBTEST o Mantel-Haenszel
No paramétricos	No uniforme	Regresión logística	Comparación por pares mediante Regresión Logística
Paramétricos	Uniforme	Test de Razón de Verosimilitud Lord	Test de Razón de Verosimilitud Lord
Paramétricos	No uniforme	Test de Razón de Verosimilitud Lord	Test de Razón de Verosimilitud Lord

7.1.1 Principales métodos no paramétricos

Test 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 del instrumento (Holland y Thayer, 1988). El valor que se obtiene de calcular este estadístico sigue una distribución $\chi 2$ con un grado de libertad. Por lo tanto, valores de

la prueba MH mayores a un valor crítico basado en la distribución $\chi 2$, nos indicarán la presencia de DIF para ese reactivo en específico. Este valor suele transformarse en la razón de probabilidades logarítmicas. 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).

Para calcular el estadístico MH, se parte de la premisa de que los ítems que están afectados por DIF tienen una correlación diferente entre los grupos que los que no están afectados (Holland y Thayer, 1988). De esta manera, se busca controlar el efecto de variables confusoras, como las diferencias en habilidad entre los grupos, para detectar si la pertenencia a un grupo está afectando la probabilidad de que un individuo conteste correctamente un ítem. Para realizar el cálculo, se deben agrupar las respuestas en una tabla de contingencia, donde se registran las respuestas correctas e incorrectas de los dos grupos. A partir de esta tabla, se estiman las proporciones de respuestas correctas para cada grupo y para cada nivel de habilidad (si se están controlando variables confusoras). Luego, se calcula un estadístico chi-cuadrado que evalúa la asociación entre la pertenencia a un grupo y la respuesta correcta a un ítem, corrigiendo por los niveles de habilidad.

Es importante tener en cuenta que el método MH tiene algunos supuestos que deben cumplirse para obtener resultados confiables. Por ejemplo, se asume que la relación entre la pertenencia a un grupo y la probabilidad de respuesta correcta es lineal, y que la proporción de respuestas correctas sigue una distribución binomial (Holland y Thayer, 1988). Además, se requiere que el tamaño de la muestra sea suficientemente grande para obtener resultados precisos y que los grupos sean equivalentes en términos de habilidad y otros factores relevantes. Si estos supuestos no se cumplen, los resultados pueden ser sesgados o poco confiables.

SIBTEST: El Test de Sesgo de Reactivos Simultáneo o SIBTEST por sus siglas en inglés calcula si la diferencia media ponderada entre el grupo focal y el grupo referencial es estadísticamente significativa. Una de las ventajas de este método es que utiliza un procedimiento

de corrección de la diferencia media utilizando un método de regresión, lo que permite ajustar el resultado ante diferencias en la distribución de habilidad entre ambos grupos. Este procedimiento de corrección ayuda a subsanar parcialmente la limitación de los modelos basados en TCT que toman el puntaje total de la prueba como único indicador del nivel de habilida de los sustentantes.

El estadístico utilizado para la prueba de hipótesis *B* sigue una distribución normal estándar asintótica. Además de utilizarse para la prueba de hipótesis, este estadístico también sirve como un indicador del tamaño del efecto de DIF. De acuerdo con Roussos y Stout (1996), cuando B es menor a .059, la presencia de DIF es insignificante, si está entre .059 y .088 es moderado, y si es mayor a .088 es DIF elevado.

Regresión logística: Se puede utilizar un modelo de regresión logística para detectar el efecto DIF. El modelo a utilizar para ello toma la pertenencia a alguno de los grupos y el puntaje total de la prueba -así como la interacción entre ambos- como parámetros para predecir la probabilidad de responder correctamente el reactivo. El efecto DIF uniforme se observa en el efecto predictivo de la pertenencia a un grupo, mientras que el DIF no uniforme se observa en el efecto de la interacción de pertenencia al grupo y puntaje total de la prueba.

Para identificar la presencia de DIF, además del uso de prueba de hipótesis mediante el contraste del valor de *p*, se utiliza el valor ΔR2. Dentro de la propuesta de Jodoin y Gierl (2001), se considera como efectos insignificantes a valores menores a .035, moderados a valores entre .036 y .07, señalando como efectos DIF grandes a cualquier valor mayor a .071. En cualquier caso, estos criterios deben ser interpretados con precaución y siempre ser acompañados por un análisis cuidadoso de los datos y la consideración de otros factores relevantes.

Es importante tener en cuenta que la regresión logística asume que la relación entre las variables predictoras y la variable respuesta es lineal en la escala logit, es decir, en la escala de las probabilidades logarítmicas. Además, el modelo también asume la ausencia de multicolinealidad,

que no haya datos faltantes o que estos estén ausentes al azar y que las observaciones sean independientes. Por lo tanto, es importante realizar un análisis cuidadoso de los supuestos antes de aplicar la regresión logística para la detección de DIF (Magis et al., 2010).

En cuanto al método de cálculo para la detección de DIF mediante regresión logística, se utiliza el valor de p para evaluar la significación estadística del efecto del grupo en la probabilidad de respuesta correcta al reactivo. Si el valor de p es menor que el nivel de significación establecido, se rechaza la hipótesis nula y se concluye que existe DIF. Sin embargo, un valor de p significativo no indica el tamaño del efecto DIF. Para evaluar el tamaño del efecto, se pueden utilizar diferentes medidas, como el índice de discriminación, el índice de correlación biserial puntual o el índice de phi. Cada medida tiene sus propias ventajas y limitaciones y es importante considerarlas en conjunto con otras fuentes de evidencia para llegar a una conclusión adecuada sobre la presencia y el tamaño del efecto DIF (Magis et al., 2010; Jodoin y Gierl, 2001).

7.1.2 Principales métodos TRI

χ2 de Lord: Este procedimiento parte de la hipótesis nula de parámetros equitativos de los reactivos para ambos grupos; al igual que el método de Mantel-Haenszel, se basa en un valor estadístico con una distribución χ2. La diferencia es que toma como valores iniciales los parámetros obtenidos de un modelo de la TRI, ya sea 1, 2 o 3 parámetros. Los parámetros obtenidos de estos modelos son transformados para obtener un valor Q que nos permite llevar a cabo el análisis de χ2 (Magis et al., 2010).

Ya que este método toma como valores iniciales los de un modelo TRI, uno de sus supuestos principales es que dicho modelo subyacente es adecuado para los datos. Esto significa que el modelo TRI elegido debe ajustarse bien y proporcionar estimaciones precisas de los parámetros del ítem y la habilidad.

Otro supuesto importante es que los grupos que se comparan deben tener una distribución similar de habilidades. Si este supuesto no se cumple, puede ser necesario utilizar técnicas de emparejamiento o estratificación para equilibrar las distribuciones de habilidades entre los grupos antes de aplicar el método de Chi-cuadrado generalizado de Lord (Magis et al., 2010).

La interpretación de este método es muy similar a la de Mantel-Haenszel, y al igual que éste, también se ha generalizado esta metodología para poder estimar el efecto DIF para más de dos grupos.

Test de Razón de Verosimilitud: Este enfoque se basa en comparar dos modelos de la TRI distintos. El primero de ellos consiste en forzar que los parámetros del reactivo para ambos grupos sean iguales; en el segundo modelo, se permite la variación de dichos parámetros entre los grupos. Nuevamente, el valor obtenido de esta comparación sigue de manera aproximada a una distribución $\chi 2$, por lo que su interpretación es muy similar a lo ya descrito. La principal diferencia está en los datos que utiliza para generar la comparativa. Dado que es necesario estimar los parámetros de dos modelos distintos de la TRI, este procedimiento es más tardado y complejo que algunos de los otros que se han descrito en la literatura (Magis et al., 2010).

7.2 Interpretación de DIF

Como se vio en la sección anterior, distintos métodos para detectar DIF ofrecen distintos tipos de valores y criterios para señalar reactivos con funcionamiento diferencial. Sin embargo, una de las cuestiones cruciales en este tipo de análisis es el qué se hace después con estos resultados, y es tema de mucha discusión científica (Cho et al., 2016). McNamara y Roever (2006) señalan que la presencia de DIF por sí misma no significa que el reactivo no sea imparcial o que deba ser eliminado. En un sentido técnico, la presencia de DIF simplemente indica qué factores irrelevantes al constructo medido están teniendo una influencia sistemática en los patrones de respuesta del reactivo.

Por ello, una vez detectados los reactivos con DIF significativo, es importante complementar el análisis con una visión más profunda de carácter cualitativo, permitiendo que estos reactivos sean revisados por expertos para llegar a un acuerdo sobre las posibles causas de DIF, para de ahí, determinar si es pertinente eliminar el reactivo, modificarlo o simplemente dejarlo tal como está y utilizar otros medios estadísticos para compensar la presencia de DIF.

Un ejemplo concreto de aplicación de análisis DIF en el contexto de exámenes de alto impacto en México fue el estudio realizado por García Medina et al. (2016) donde analizaron el funcionamiento diferencial en reactivos pertenecientes al examen EXCALE en su componente de matemáticas para tercero de secundaria, comparando a los sustentantes a partir de su sexo y nivel socioeconómico. Mediante el uso del método de diferencial de logits a partir de un modelo Rasch, reportan que ninguno de los 100 reactivos presenta DIF por sexo, pero 18 presentan DIF severo por estatus socioeconómico.

De manera complementaria, Mendoza Vega y Corona Burch (2018) evaluaron también la presencia de reactivos con funcionamiento diferencial para la escala de Español del examen Excale, pero en este caso, tomando como grupo focal a aquellos sustentantes que hablasen una lengua indígena. Estos autores contrastaron tres metodologías para detectar DIF distintas: el método de Mantel-Haenszel, la Ji cuadrada de Lord y el área exacta de Raju. Los resultados fueron un total de 49 de 169 reactivos señalados con presencia de DIF por al menos uno de los métodos. Una práctica importante a destacar que siguieron los autores de este artículo fue el análisis cualitativo posterior para identificar las posibles causas del DIF, desde técnicas como minería de textos hasta el análisis cualitativo de las especificaciones.

7.3 Validez y sesgo en la evaluación

Si bien en la sección anterior se presenta una serie de alternativas para detectar DIF, es importante destacar que estas no son herramientas aisladas que debamos tomar a conveniencia

según nos resulte más sencillo llevar a cabo un análisis u otro. Si bien existen estudios que comparan la efectividad de algunos de estos modelos en contextos específicos (Apinyapibal et al., 2015; Kristjansson et al., 2005), la realidad es que es necesario considerar estos análisis dentro del marco general que se esté utilizando para la evaluación de la validez de las interpretaciones y usos del instrumento a valorar. De ahí que los análisis del funcionamiento diferencial de los reactivos, como una evidencia de imparcialidad, se vuelven parte de los procesos de validación antes mencionados.

Dado que los análisis DIF son parte de un proceso más amplio, es importante insertarlos dentro de un mismo marco teórico y proceso metodológico que genere un nivel de concordancia suficiente entre las distintas evidencias de validez presentadas. En un ejemplo concreto, si la validez con respecto al correcto funcionamiento de los reactivos y a su dimensionalidad se basó en métodos de la Teoría de Respuesta al Ítem, es importante que los análisis DIF sigan los mismos supuestos y estén enmarcados dentro de la misma teoría y principios. Si, por ejemplo, el análisis de la dimensionalidad mostrara que el instrumento es multidimensional, pero el análisis DIF se hiciera bajo el supuesto de que se trata de un instrumento unidimensional, se estarían generando evidencias contradictorias.

8. Teoría de Respuesta la Ítem

La TRI busca conseguir mediciones que no dependan específicamente ni de las personas evaluadas ni de los reactivos utilizados. Para ello, se generaron modelos matemáticos que asumen que la probabilidad de que una persona emita cierta respuesta ante un reactivo puede describirse a partir de la ubicación de la persona en el rasgo o dimensión que se está midiendo y de una o más características del reactivo (Bock y Moustaki, 2006).

Al hablar de probabilidad de acierto de los reactivos, dentro de los modelos TRI se puede visualizar mediante la *Curva Característica del Ítem (CCI).* Esta curva es una función logística de la

forma $f(x)\, {e^x}/{1 + e^x}$. La *CCI* es una representación gráfica del funcionamiento de cada reactivo (Leenen, 2014). En el eje de las X se representa el valor de θ (nivel de habilidad), mientras que en el eje de las Y se representa la probabilidad de cumplir con el criterio específico que mide el reactivo. Dado que son reactivos dicotómicos, este eje presenta únicamente valores desde 0 hasta 1. Para obtener la representación específica de la curva de cada reactivo, se calculan los siguientes parámetros:

- Nivel de habilidad θ: El valor de θ, en el contexto educativo, suele representar el nivel de habilidad general del sustentante, este valor se obtiene mediante cálculos matemáticos complejos que consideran el desempeño general de cada sustentante en todos los reactivos considerados unidimiensionales.

- Parámetro de dificultad (*b*): Este parámetro, dentro del modelo matemático, describe el nivel necesario que cada observación requiere tener en el rasgo o dimensión que mide el instrumento para obtener una probabilidad mayor a .5 de acertar en el reactivo evaluado. En ese sentido, el valor estaría dado por el punto de intersección en el eje de las X con el valor de .5 en el eje de las Y dentro de la *CCI*. De manera general, se suele utilizar un criterio de aceptabilidad de los reactivos cuyo valor del parámetro *b* oscile entre -2.5 y 2.5; dado que estos números corresponden a valores de θ, valores negativos implican que sustentantes con menor nivel de habilidad tienen una alta probabilidad de acierto, mientras que valores positivos del parámetro *b* implican que solo sustentantes con niveles de habilidad alto tienen una probabilidad elevada de acertar.

- Parámetro de discriminación (*a*): Se le denomina así dado que permite identificar el cambio en la probabilidad de acierto a lo largo del nivel habilidad, lo que da indicios de qué tan bueno es el reactivo para separar a los sustentantes con mayor o menor nivel de habilidad. Sin embargo, en términos matemáticos, este parámetro indica la magnitud de cambio de la probabilidad de acierto a medida que el nivel de θ cambia; su valor es proporcional a la pendiente

de la curva dentro de la *CCI*. Dado que lo que se espera es que los reactivos sean capaces de discriminar a quienes tienen mayor habilidad, se suele utilizar como criterio para determinar la calidad de los reactivos, valores mayores a .45.

- Parámetro de pseudo-adivinación (*c*): Este parámetro suele interpretarse como la probabilidad que tienen de acertar en el reactivo aquellos sujetos que desconocen la respuesta correcta. En términos matemáticos, este parámetro representa el valor de la probabilidad de acierto cuando el nivel de θ tiende a su valor mínimo -∞.

8.1 Distintos modelos TRI

- Modelo de 1 parámetro: El modelo de Rasch, si bien presenta algunas características particulares en cuanto a estimación de parámetros y supuestos, es en esencia un modelo de 1 parámetro, puesto que iguala el parámetro *a* y asume que el *c* es igual a 0 en todos los reactivos del instrumento, la única variabilidad que encuentra es en términos del parámetro *b*, obteniendo reactivos con mayor o menor nivel de dificultad (Leenen, 2014).

- Modelo de 2 parámetros: Permite la variación en el parámetro *a*, identificando que cada reactivo puede presentar una pendiente de probabilidad diferente. Aun así, sigue asumiendo que el parámetro *c* es igual a 0 para todos los reactivos.

- Modelo de 3 parámetros: Además de la variabilidad en los parámetros *a* y *b*, este modelo permite que los reactivos tengan un valor inicial de la asíntota inferior distinto a 0.

8.2 Supuestos de los modelos TRI

Los modelos de la TRI recién presentados parten de algunos supuestos que necesitan cumplirse para poder utilizarlos. Los supuestos principales que hay que vigilar dentro de la TRI se destacan cuatro:

Monotonicidad: El supuesto de monotonicidad asume que la relación entre la habilidad y la probabilidad de acierto se puede modelar mediante una función logística, en otra palabas, indica que la probabilidad de dar la respuesta correcta a un reactivo aumenta monotónicamente en la medida en que se incrementa el nivel de habilidad θ. Esto significa que si dos individuos tienen diferentes niveles de habilidad, el que tenga un nivel más alto tiene una mayor probabilidad de responder correctamente a cualquier ítem que el que tiene un nivel más bajo (Bock y Moustaki, 2006).

Invarianza en los parámetros de los reactivos y en el nivel de habilidad: El supuesto de invarianza en la teoría de respuesta al ítem (TRI) establece que la relación entre las respuestas de los participantes y los parámetros de los ítems es la misma para todos los grupos de participantes. En otras palabras, se espera que las características psicométricas de los ítems, como la dificultad y la discriminación, sean consistentes para diferentes grupos de participantes, como diferentes edades, géneros o culturas.

El supuesto de invarianza es importante en la TRI porque la medición de habilidades o rasgos se basa en la comparación de las respuestas de diferentes grupos de participantes. Si los parámetros de los ítems varían entre los grupos, entonces las comparaciones de habilidades o rasgos entre los grupos no serán precisas o justas. Por lo tanto, la invarianza es un requisito importante para la validez y la fiabilidad de la medición en la TRI. Se destacan tres tipos importantes de invarianza: la de dificultad y la de discriminación se refieren a la consistencia de estos parámetros dentro del modelo TRI, mientras que la de grupo se refiere a la consistencia de la dificultad y discriminación de los ítems en diferentes subgrupos dentro de un grupo más grande de participantes (Bock y Moustaki, 2006).

La diferencia entre la invarianza y el DIF es que la primera se refiere a la consistencia de la medida de la habilidad del examinado en diferentes grupos, mientras que el DIF se refiere a la

consistencia de la medida del ítem en diferentes grupos de examinados. La invarianza se enfoca en asegurar que las medidas de habilidad sean comparables entre diferentes grupos, mientras que el DIF se enfoca en asegurar que el rendimiento en el ítem sea comparable entre éstos.

La independencia local o independencia condicional indica que la probabilidad condicional de observar un patrón de respuesta dado un valor específico del rasgo o habilidad latente es igual al producto de las probabilidades condicionales de los reactivos (Liu y Maydeu-Olivares, 2013). En otras palabras, la probabilidad de responder correctamente a algún reactivo en particular, una vez considerado el valor de θ no se ve influenciada por la respuesta a otros reactivos del instrumento. Si bien se podría esperar que exista una correlación entre las respuestas a distintos reactivos del mismo instrumento, dado que suponemos que todos los reactivos miden el mismo constructo, se esperaría que cada reactivo fuera independiente de los demás y que lo único que generara una correlación entre ellos fuera el nivel de habilidad.

A partir de esto, uno de los métodos más utilizados para valorar el cumplimiento de este supuesto es el estadístico Q3 de Yen (1993), que corresponde a una correlación producto-momento de Pearson entre los residuales de cada reactivo dentro del modelo de la TRI, obtenidos de la diferencia entre las respuestas observadas y las predichas por el modelo. Dado que se trata de una correlación, este estadístico puede tomar valores de -1 a 1, siendo valores más cercanos a 0 indicadores de independencia local. Existen algunas sugerencias de punto de corte para señalar pares de reactivos con dependencia local, por ejemplo, el propio Chen y Thissen (1997a) recomendó señalar valores mayores a .2 como signo de dependencia local; pese a ello, dado que el comportamiento de este estadístico se ve influenciado por el número de reactivos y el tamaño de muestra, no existe consenso en cuanto un punto de corte adecuado para señalar dependencia local.

Una segunda alternativa para valorar el supuesto de independencia local es el estadístico de la $\chi 2$ de Pearson para Dependencia Local $\chi 2\_LD$ propuesto por Chen y Thissen (1997) para

comparar los valores predichos por el modelo con los valores reales encontrados en los datos para cada par de reactivos. Este estadístico resulta ser de mayor utilidad dado que permite la contrastación de hipótesis de una forma más directa mediante el contraste de los valores *p* a partir de la distribución $\chi 2$.

Unidimensionalidad: Por último, otro supuesto de suma importancia para los modelos TRI, y que se relaciona de manera directa con el anterior es el referente a la dimensionalidad del instrumento. Reise et al. (2014) mencionan que: "un set de respuestas a reactivos es unidimensional si y solo si la matriz de respuestas es localmente independiente después de eliminar un único factor latente común".

Pese a ello, tal como mencionan Reise et al. (2014), los datos de respuesta a reactivos son raramente unidimensionales en un sentido estricto, por lo que suele ser necesario decidir si son "suficientemente unidimensionales" para los modelos TRI.

Un problema importante que surge de esto es que, si datos multidimensionales se forzan en modelos unidimensionales, la estimación de parámetros de los reactivos se puede ver distorsionada de forma severa, por lo que cualquier uso o interpretación dado a partir de dicho modelo debe ser seriamente cuestionado.

8.3 Modelos multidimensionales

Ahora bien, el supuesto de unidimensionalidad es difícilmente alcanzable en la mayoría de los constructos. Desde el siglo pasado, se constató la necesidad de considerar la dimensionalidad de las pruebas al momento de buscar la detección de DIF (Mazor et al., 1998), dado que no considerarlo puede incrementar considerablemente el nivel de error, particularmente en cuanto a falsos positivos (reactivos señalados con DIF que realmente no lo presentan). Para ilustrar el por qué de este punto, es necesario hacer una breve recopilación respecto a los distintos tipos de modelos basados en Teoría de Respuesta al Ítem Multidimensional (MIRT por sus siglas en inglés).

Los modelos MIRT pueden entenderse como una generalización de los modelos unidimensionales que permite ir más allá del supuesto de unidmensionalidad. En ese sentido, los modelos de Rasch, de 2 o de 3 parámetros son fácilmente generalizables para ser multidimensionales. Así, por ejemplo, en el modelo de dos parámetros (2PL), la probabilidad de respuesta correcta a determinado reactivo es una función del nivel de habilidad del sustentante y de los parámetros *a* y *b*; por otro lado, en un modelo MIRT, la probabilidad de respuesta correcta se modela como una función de los parámetros *a* y *b* y de múltiples dimensiones de habilidades.

Dependiendo de las relaciones que haya entre estas dimensiones y las probabildiades de acierto de los reactivos, se desprenden distintos tipos de modelos MIRT. De acuerdo con Hartig y Höhler (2009), son tres las cuestiones fundamentales a tomar en cuenta para clasificar los distintos tipos de modelos MIRT.

Multidimensionalidad entre o dentro de los ítems: Si bien un modelo MIRT implica la existencia de más de una dimensión necesaria para la resolución de la prueba, a nivel de reactivo, la relación de estas dimensiones puede ser muy distinta. Cuando la multidimensionalidad se da *entre* los reactivos, significa que la probabilidad de responder correctamente un reactivo dado está en función de una sola de las múltiples dimensiones incluidas en la prueba. En otras palabras, algunos reactivos evalúan específicamente una de las dimensiones mientras que otros reactivos evalúan alguna otra por separado.

Por otro lado, cuando la multidimensionalidad se da *dentro* de los reactivos, entonces la probabilidad de respuesta correcta de un reactivo dado estará en función de más de una dimensión de habilidad de manera simultánea. Así, el reactivo requeriría de más de una habilidad para su resolución.

Modelos MIRT compensatorios o no compensatorios: A partir del segundo caso donde la multidimensionalidad se encuentra dentro de los reactivos, es necesario hacer una segunda

distinción: La relación existente entre estas múltiples dimensiones al momento de calcular la probabilidad de una respuesta correcta. En el caso de los modelos compensatorios, un nivel bajo en una de las habilidades podría ser compensado por un nivel alto en otra de las dimensiones de habilidad al momento de determinar la probabilidad de acierto. En otras palabras, este tipo de modelo asume que hay más de una forma de llegar a la respuesta correcta de un reactivo dado, ya sea haciendo uso de una sola o más de las habilidades modeladas como necesarias para dicho reactivo.

En el caso de los modelos no compensatorios, se asume que, para tener una alta probabilidad de respuesta correcta en un reactivo dado, es necesario tener un nivel alto de habilidad en todas las dimensiones involucradas en este reactivo. Este podría ser el caso de reactivos que impliquen tareas secuenciales, donde en la primera tarea se requiera usar una de las habilidades y en la segunda tarea se requiera de otra. Así, aunque se tenga un alto nivel de habilidad en la segunda, si no es posible resolver la primera parte debido a un bajo nivel de habilidad en la primera tarea, la probabilidad de respuesta correcta será baja.

Número de dimensiones: Por último, Hartig y Höhler (2009) señalan como un apartado crucial el referente a la determinación del número adecuado de dimensiones a incluir dentro de nuestro modelo. Para ello, es importante partir del propósito ya descrito que hay detrás de los modelos multidimensionales: evaluar las distintas habilidades requeridas para un buen desempeño dentro del área que evalúa la prueba. A partir de este propósito, podemos ver que el punto clave es determinar a qué nos referimos con estas "distintas habilidades". Ya desde hace tiempo, Briggs y Wilson (2003) señalaban que la gran mayoría de áreas que se evalúan dentro del contexto de la educación pueden ser tan complejas que bien se podría desprender tantas o más dimensiones de habilidad como reactivos dentro de la prueba. Si bien este es el caso, desarrollar modelos a ese nivel de especificidad resultaría poco práctico, no solo en cuanto a la generación del modelo a nivel matemático, sino también a nivel de su interpretabilidad.

Por ello, determinar el número de dimensiones a incluir implica determinar en qué nivel de complejidad se ha de ubicar el modelo. Para Hartig y Höhler (2009), los distintos dominios de contenido se pueden conceptualizar de forma jerárquica en términos de su nivel de generalidad o especificidad. Por ejemplo, asumiendo que una prueba mide un dominio de conocimientos específicos como escritura, estaríamos partiendo del nivel jerárquico de generalidad más alto, donde asumimos que todos los reactivos de la prueba miden escritura. Si fuéramos a un nivel más profundo, podríamos determinar que para evaluar escritura como lo plantea la prueba, se puede dividir en las dimensiones de redacción y de gramática, las cuáles a su vez se podrían nuevamente desglosar hasta alcanzar un nivel de especificidad muy elevado.

Entonces, el tema central sería encontrar un balance adecuado entre generalidad y especificidad. Dentro del contexto psicométrico se utiliza el término de parsimonia para describir justo eso: encontrar el menor número de dimensiones posibles que sean tanto significativas a nivel estadístico dentro del modelo como a nivel práctico y de interpretación (Briggs y Wilson, 2003). Esta determinación dependerá también del marco teórico y del propósito de la prueba desde el que se parta.

8.4 Métodos para seleccionar el mejor modelo

Si bien en evaluación solemos partir de una tabla de especificaciones que explicita qué dominios y subdominios evalúa cada reactivo, es importante constatar que los reactivos y la prueba en general verdaderamente se comporten como se hipotetizó que lo harían. Para ello, es importante modelar los datos utilizando los parámetros y dimensiones que mejor se ajusten a lo hipotetizado, para posteriormente contrastarlo con otros modelos plausibles, para determinar cuál de esos modelos explica de mejor manera lo observado en los datos.

8.4.1 Bondad de ajuste de los modelos

La bondad de ajuste de un modelo describe qué tan bien el modelo se ajusta a lo observado en los datos. Para medir la bondad de ajuste existen dos opciones principales: los índices de bondad

de ajuste y los estadísticos de bondad de ajuste. Los primeros sirven para resumir las posibles discrepancias entre los datos esperados por un modelo estadístico y los datos observados; los segundos corresponde a índices de bondad de ajuste con distribuciones de muestreo conocidas y que, por lo tanto, son utilizados en la comprobación de hipótesis (Maydeu-Olivares, 2014).

-CFI: El índice de ajuste comparativo (CFI por sus siglas en inglés) es un índice de la bondad de ajuste del modelo propuesto que se obtiene de compararlo con un modelo base. Este índice va de 0 a 1, y se suelen considerar como aceptables valores mayores a .96 (Hu y Bentler, 1999), aunque investigaciones recientes muestran que este valor de punto de corte debería determinarse a partir de las condiciones específicas del modelo y de la muestra (McNeish y Wolf, 2023).

-RMSEA: La media cuadrática del error de aproximación (RMSEA por sus siglas en inglés) es una medida de ajuste absoluta, dado que establece la distancia entre el nivel de ajuste de los valores predichos por el modelo y los valores reales en términos de varianza. Este estadístico es normalmente utilizado dentro del contexto de Modelos de Ecuaciones Estructurales. En el contexto de la TRI, Maydeu-Olivares (2014) propuso un RMSEA basado en una muestra bivariada. Dado que es una medida de error, entre más pequeño sea el valor de RMSEA mayor se considera el nivel de ajuste. El propio Maydeu-Olivares (2014) sugirió un punto de corte de valores menores a .05.

8.4.2 Test de razón de verosimilitud

El test de razón de verosimilitud es una técnica estadística que se utiliza para comparar dos modelos de la teoría de respuesta al ítem (TRI) y determinar si uno de los modelos es significativamente mejor que el otro en términos de ajuste a los datos. Para realizar el test de razón de verosimilitud, se ajusta un modelo nulo (modelo simplificado) y un modelo alternativo (modelo completo) a los mismos datos de prueba. El modelo nulo es una versión más simple del modelo alternativo y se utiliza como punto de referencia para comparar el modelo completo. Por ejemplo, el modelo nulo puede ser un modelo con parámetros de dificultad iguales para todos los ítems,

mientras que el modelo alternativo puede ser un modelo con parámetros de dificultad diferentes para cada ítem.

Una vez que se han ajustado los dos modelos, se calcula la diferencia en la verosimilitud entre ellos. Esta diferencia se llama estadística de prueba y se puede usar para calcular un valor p que indica la probabilidad de que los datos observados sean consistentes con el modelo nulo. Si el valor *p* es menor que un nivel de significancia previamente establecido (por ejemplo, 0,05), se puede rechazar el modelo nulo y aceptar el modelo alternativo como el modelo más adecuado para los datos. Esto significa que el modelo alternativo proporciona un mejor ajuste a los datos y es más preciso para describir la relación entre las respuestas de los participantes y las características de los ítems.

8.4.3 M2

El estadístico M2 es una prueba estadística del nivel de ajuste del modelo basado en los residuales que se utiliza de manera especial en datos binarios como los presentes en este estudio (Maydeu-Olivares y Joe, 2006). Esta prueba parte de una hipótesis nula de igualdad en términos de probabilidades calculadas a partir del modelo seleccionado y del nivel de theta calculado a partir de éste (la cantidad de parámetros que utiliza el modelo es lo que otorga los grados de libertad (*df*) para la prueba de significancia). Por ello, en este estadístico, que busca ver el nivel de ajuste del modelo, se esperaría que la prueba no resultara significativa. Valores de p menores a .05 indicarían un mal ajuste del modelo a los datos. De igual forma, en una comparativa directa entre dos modelos, el valor de M2 que sea menor es indicio de un mejor nivel de ajuste (Maydeu-Olivares y Joe, 2006).

En complemento al valor de M2, se puede utilizar la estadística AIC (Akaike's Information Criterion, por sus siglas en inglés) o BIC (Bayesian Information Criterion) para determinar cuál de los dos modelos es mejor. La estadística AIC y BIC son medidas de la calidad de ajuste de un modelo a los datos, que penalizan el modelo por su complejidad (número de parámetros). La idea es

encontrar el modelo que tenga un buen ajuste a los datos con el menor número posible de parámetros (Cavanaugh y Neath, 2019; Neath y Cavanaugh, 2012).

La estadística AIC se calcula como AIC = -2 log(L) + 2k, donde L es la verosimilitud del modelo y k es el número de parámetros del modelo (Cavanaugh y Neath, 2019). La estadística BIC se calcula de manera similar, pero penaliza más la complejidad del modelo, y se define como BIC = -2 log(L) + log(n)k, donde n es el tamaño de la muestra (Neath y Cavanaugh, 2012). En general, se prefiere el modelo con un valor de AIC o BIC más bajo, lo que indica un mejor ajuste con una menor complejidad. Es importante tener en cuenta que la comparación de modelos con estas estadísticas solo debe realizarse si ambos modelos se ajustan a los mismos datos. Si se utilizan diferentes conjuntos de datos para ajustar cada modelo, la comparación no es válida

III. CONTEXTO DEL POSGRADO Y LOS EXÁMENES DE ADMISIÓN

1. El posgrado en la IES

La Institución de Educación Superior (IES) con la que se trabajó es una de las universidades más importantes del país. A nivel de posgrados, cuenta con cerca de 14 mil alumnos entre maestrías y doctorados. Muchos de sus programas están acreditados con niveles altos dentro del Padrón Nacional de Posgrados de Calidad (PNPC) del Consejo Nacional de Ciencia y Tecnología. En total, la IES cuenta con 42 programas de los que se derivan 57 planes de estudio distintos de maestría y 38 de doctorado. Estos programas están a su vez divididos en cuatro áreas de conocimiento (Torres Labansat, 2022):

- I: Ciencias Físico-Matemáticas y de las Ingenierías (CFMI)

- II: Ciencias Biológicas y de la Salud (CBQS)

- III: Ciencias Sociales (CS)

- IV: Humanidades y de las Artes

2. Ingreso al posgrado

El proceso de admisión al posgrado de esta IES es complejo y masivo. En 2022, a través de dos periodos distintos de convocatoria para la admisión, se contablizó un total de 18,949 registros de aspirantes, de los cuáles, solamente 4275 fueron admitidos al posgrado (Torres Labansat, 2022).

La complejidad de este proceso incrementa al considerar que el proceso de admisión no se encuentra estandarizado, pues es responsabilidad de la coordinación de cada programa de estuios definir los procesos de admisión correspondientes. Inclusive, dentro de un mismo programa, aunque se cuente con un examen general de conocimientos estandarizado, otras fases del proceso

de admisión se dejan a criterio de las coordinaciones de cada uno de los planes de estudio que componen a dicho programa.

Por ello, con el interés de generar una herramienta útil 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.

3. Propuesta de examen general

En 2017 la Red Colaborativa 1 de la Comisión Permanente del Posgrado del Consejo de Evaluación Educativa de la IES concluyó que la evaluación de habilidades generales en los aspirantes era necesaria para la gran mayoría de los programas de posgrado. Los componentes de esta prueba deberían ser los ejes transversales para una evaluación complementaria de ingreso a estos programas 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 esta Comisión consideraron pertinente realizar un estudio que valorara la eficacia de la Prueba de Habilidades Verbales, diseñada por la Dirección de Evaluación Educativa (DEE) como un instrumento para ser aplicado a los aspirantes a ingresar a alguno de los programas de posgrado independientemente del área de conocimiento.

Dado el alto impacto que puede llegar a tener el uso previsto de estos exámenes, es necesario presentar evidencias de sus diversas interpretaciones. Al momento, ya se cuenta con evidencias preliminares de su buen desempeño psicométrico en términos de dificultad y discriminación de los reactivos a partir de modelos unidimensionales. Además de esto, se cuenta

con evidencia de su capacidad predictiva respecto al desempeño de los alumnos admitidos en uno de los programas de posgrado de la IES (Martínez González et al., 2017).

El análisis DIF ha ido cobrando mayor atención y relevancia a través de los años, convirtiéndose en la actualidad en una importante evidencia de validez tanto en términos de imparcialidad como de estructura interna de los instrumentos (AERA et al., 2014). En especial, en exámenes de alto impacto, cobra mayor relevancia la generación de este tipo de evidencias de validez, dados los posibles efectos positivos o negativos que puede tener en la vida de las personas.

El presente estudio tiene como propósito generar evidencias de validez de las posibles interpretaciones de imparcialidad en el proceso de admisión en cuatro áreas de posgrado y analizar la estructura interna de los exámenes de comprensión lectora, redacción y gramática generados por una dependencia de evaluación educativa en una institución de educación superior. El estudio incluye el análisis de la dimensionalidad de los exámenes y el análisis DIF utilizando el modelo que mejor ajuste a los datos. Así mismo, se pretende comparar los resultados de distintos métodos para analizar DIF, a fin de valorar su nivel de concordancia y discutir las posibles implicaciones a partir de los supuestos de cada uno de los análisis propuestos.

IV. MÉTODO

1. Justificación

Las evidencias de validez de usos e interpretaciones de pruebas de alto impacto requieren contar con un cierto nivel de coherencia, por lo que es necesario utilizar técnicas DIF que vayan acordes a los métodos utilizados en la generación de otro tipo de evidencias de validez. Dado que los análisis DIF parten de evaluar diferencias en la probabilidad de acierto en los reactivos, mediando el nivel de habilidad, es necesario valorar la dimensionalidad de los instrumentos para constatar si se trata de una única habilidad siendo evaluada por el examen, especialmente tratándose de habilidades complejas como lo son las habilidades verbales de comprensión lectora, redacción y gramática (Thirakunkovit, 2018). Por ello, se plantean las siguientes preguntas de investigación:

- ¿Es válido interpretar los resultados del examen como imparciales entre las distintas áreas del posgrado?
- ¿Qué reactivos del examen de comprensión lectora, redacción y gramática presentan un efecto DIF significativo?
- ¿Qué metodología es la más adecuada para obtener DIF en este examen y en muestras específicas?

2. Objetivos

General

Generar evidencias de validez de la interpretación de que los resultados de los exámenes de Comprensión Lectora, Redacción y Gramática son imparciales entre las distintas áreas de posgrado. Esto mediante el uso de técnicas para detectar DIF.

Específicos

Determinar el modelo estadístico y la metodología de DIF más adecuadas para las condiciones específicas de cada uno de los exámenes.

Comparar distintas metodologías para la obtención del efecto DIF en los reactivos del examen para identificar los métodos más adecuados, precisos y eficaces en la detección de DIF en este contexto

Identificar los reactivos de ambos exámenes que presenten efecto DIF significativo.

Presentar una guía para la obtención del efecto DIF mediante distintas metodologías a través del uso del software estadístico R.

3. Hipótesis

- Los reactivos de los exámenes de comprensión lectora, redacción y gramática no presentarán un funcionamiento diferencial significativo entre las cuatro áreas del posgrado de la UNAM.

- No habrá diferencias entre los resultados de las distintas metodologías utilizadas.

- Ambos exámenes presentarán una estructura unidimensional

4. Participantes

Con la finalidad de contar con una muestra proveniente de las cuatro áreas de conocimiento del Posgrado de la IES, se aplicó el instrumento a distintos programas de posgrado en diferentes fases, desde el ciclo 2019-1 al 2020-2. La información de las áreas y posgrados participantes se puede observar en la Tabla 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. Todos los posgrados incluidos en esta aplicación participaron de manera voluntaria con la previa aprobación de sus respectivos coordinadores.

Tabla 2.

Distribución de examinados en áreas, posgrados y periodos académicos

Área	Posgrados	Periodos	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

5. Instrumentos

El diseño, construcción y aplicación de los exámenes fue llevado a cabo por la Dirección de Evaluación Educativa (DEE) de la IES para la que se diseñó el examen, y todos los procesos aquí listados fueron llevados a cabo a partir de los criterios y metodologías internas de dicho departamento.

Para la construcción de los dos instrumentos, una comisión externa de académicos expertos en español elaboró las tablas de especificaciones de los dos exámenes que comprende la Prueba de Habilidades Verbales: 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 del DEE, los profesores elaboraron y validaron los reactivos de la prueba. Todos los reactivos diseñados fueron de opción múltiple con cuatro opciones de respuesta. La tabla de especificaciones para el examen de comprensión lectora consideró una única dimensión, mientras que la del examen de redacción y gramática consideró cuatro dimensiones: reglas gramaticales, redacción, vocabulario y ortografía.

Para ensamblar la prueba solo se consideraron los reactivos con indicadores estadísticos deseables de acuerdo con en la Teoría Clásica de los Tests y los modelos de uno y dos parámetros de la Teoría de Respuesta al Ítem. Cabe señalar que estos reactivos fueron aplicados a diferentes poblaciones de aspirantes a programas de posgrados de la IES, y posteriormente se calibraron, a fin de contar con sus parámetros 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 Teoría Clásica de los Tests (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. También se cuidó el comportamiento de las opciones de respuesta, es decir, se seleccionaron reactivos donde la opción correcta fuera elegida por una mayor proporción del grupo de alto desempeño (examinados con mayor número de aciertos en el componente de la prueba) en comparación con el de bajo desempeño (examinados con menos aciertos en la prueba); en las opciones incorrectas se verificó que las eligiera una mayor proporción del grupo de bajo desempeño.

En el índice de discriminación se consideraron reactivos que permitieran distinguir entre los grupos de alto y bajo desempeño. 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. 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 Teoría de Respuesta al Ítem se consideró 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.

Después de la calibración, la prueba se conformó por 31 reactivos de opción múltiple con cuatro opciones de respuesta; 11 reactivos para el examen de Comprensión lectora y 20 reactivos para el examen de Redacción y gramática; estos 31 reactivos cumplieron los requisitos anteriormente mencionados.

Se ensamblaron dos versiones de la prueba integradas con los mismos reactivos, las cuales sólo difieren en el orden en el que se encuentran presentados los exámenes,

6. Procedimiento

La Dirección de Evaluación Educativa (DEE) de la IES responsable del examen, 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 del 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 del 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.

7. Análisis estadísticos

Se comenzó el análisis por mostrar de manera descriptiva las puntuaciones obtenidas por cada una de las áreas de conocimiento. Se realizaron análisis descriptivos a nivel del reactivo y de puntaje total, comparando el porcentaje de respuestas correctas por reactivo por área de conocimiento, y también se observaron diferencias en la distribución de puntajes totales entre las cuatro áreas. Además, se analizó de manera general la distribución de puntajes para valorar la normalidad de la muestra.

Para comenzar con el análisis de la presencia de DIF en los reactivos de ambos exámenes, se utilizaron métodos convencionales de obtención de DIF tanto desde la TRI como a partir de métodos no paramétricos. Esto con la finalidad de poder observar si existían diferencias en cuanto a los reactivos señalados con funcionamiento diferencial entre los distintos métodos, para posteriormente comparar estos resultados con los obtenidos mediante las técnicas más robustas. Tal como señala Hambleton (2006), presentar evidencias de validez desde metodologías diversas que a su vez se basan en supuestos y procedimientos distintos, permite fortalecer las evidencias de validez.

Los métodos que se utilizaron para este análisis comparativo fueron el de Mantel-Haenszel y el de regresión logística dentro de los métodos no paramétricos para evaluar DIF uniforme y no uniforme en comparación de dos grupos. En cuanto a métodos paramétricos, se utilizaron el de diferencia de Logits a partir de un modelo Rasch y el método de χ^2 de Lord, ambos casos para medir DIF uniforme.

Dado que estos métodos son más informativos al comparar directamente dos grupos (grupo focal y grupo referencia), se procedió a realizar el análisis desde dos perspectivas distintas. Por un lado, se analizó la presencia de DIF en el contraste directo de área contra área; al tener cuatro distintas, se realizaron seis análisis para poder valorar la interacción entre cada par de áreas. De

manera adicional, se realizaró el contraste de las áreas CFMI y CBQS del posgrado contra las CS y HA.

Esta combinación se realizó por dos motivos principales: Dentro del propio DEE, se partía como una hipótesis inicial que las diferencias se darían principalmente entre las áreas CFMI y CBQS y las áreas CS y HA. Por otro lado, dado el tamaño de muestra de las áreas CFMI y HA, puede resultar más conveniente un análisis de dos muestras más grandes, especialmente en los análisis más robustos como los basados en la TRI. Se contrastaron las diferencias entre ambos análisis, así como las diferencias con otros métodos para estimar DIF, tanto con dos como con cuatro grupos.

Para llevar a cabo los análisis DIF en ambos exámenes a partir de la TRI, se comenzó por seleccionar el modelo más parsimonioso con un buen ajuste para cada uno de los exámenes, vigilando aspectos como la cantidad de parámetros de los reactivos a considerar dentro de los modelos, la dimensionalidad de ambos instrumentos y el cumplimiento de supuestos como el de independencia local. Para determinarlo, se comparó el nivel de ajuste de varios modelos.

En un primer momento, se compararon únicamente modelos unidimensionales para evaluar si un modelo de uno, dos o tres parámetros era el más indicado para representar los datos obtenidos. Para poder compararlos, se tomaron como criterios los estadísticos M2, CFI y RMSEA ya descritos previamente. Adicionalmente, se utilizó la prueba de razón de verosimilitud para comparar de manera directa dos modelos.

Posteriormente, tras elegir el modelo unidimensional que mejor representaba los datos, se procedió a comparar su nivel de ajuste con el de modelos multidimensionales con la misma cantidad de parámetros, tanto modelos provenientes de la estructura factorial propuesta como modelos exploratorios; la comparación de estos modelos unidimiensionales y multidimensionales se llevo a cabo con los mismos criterios.

Una vez seleccionado el mejor modelo en términos de parámetros y dimensionalidad, se procedió a verificar el supuesto de independencia local de ambos exámenes bajo los modelos seleccionados. Para ello, se utilizó el estadístico Q3, tomando como criterio para determinar la presencia de dependencia local valores mayores a .2, como es recomendado por De Ayala (2013). Dado que este punto de corte puede contener cierto nivel de sesgo, se complementó el análisis mediante el estadístico χ^2LD propuesto por Chen y Thissen (1997). Se determinó que un reactivo no presenta dependencia local solamente si no es señalado por ninguna de los dos estadísticos.

Tras verificar el cumplimiento de los supuestos de la TRI y seleccionar el mejor modelo, se procedió a analizar la presencia de DIF en ambos exámenes. Para ello, se utilizó el método del test de razón de verosimilitud. El análisis se llevó a cabo con la estrategia de DIF-libre-luego-DIF (DIF-Free-Then-DIF, Wang et al., 2012), tal como se describió con anterioridad.

Para estimar los modelos TRI correspondiente para llevar a cabo el análisis, se utilizó la función *multipleGroup* del paquete *mirt* (Chalmers, 2012). Esta función corre un análisis de grupos múltiples de máxima verosimilitud e información completa, permitiendo así estimar parámetros específicos para cada uno de los grupos a analizar. Para la estimación de estos modelos, se utilizó el algoritmo EM de esperanza-maximización.

Una vez elaborados estos análisis, se procedió a replicarlos con solamente dos grupos como se describió en la sección anterior, combinando como un solo grupo a los participantes de las áreas de conocimiento CFMI y CBQS, y como otro grupo a los pertenecientes a las áreas CS y HA.

7.1 Software utilizado

Todos los análisis presentados fueron realizados en el software de análisis estadístico R. Para los análisis TRI se utilizó el paquete *mirt* (Chalmers, 2012) . En cuanto a los análisis DIF no paramétricos, se utilizó el paquete *difR* (Magis et al., 2010). Para el manejo y limpieza de los datos, así como la generación de gráficas, se utilizó la familia de paquetes del *tidyverse* (Wickham et al.,

2019). Para visualizar con más detalle los análisis realizados y el código utilizado para obtenerlos, se refiere al lector al anexo 1. Adicionalmente, para una descripción más detallada a modo de tutorial para la realización de este tipo de análisis, se refiere al lector al siguiente enlace: https://davindiaz.netlify.app/posts/sintaxis_dif_tesis_unam/

V. RESULTADOS

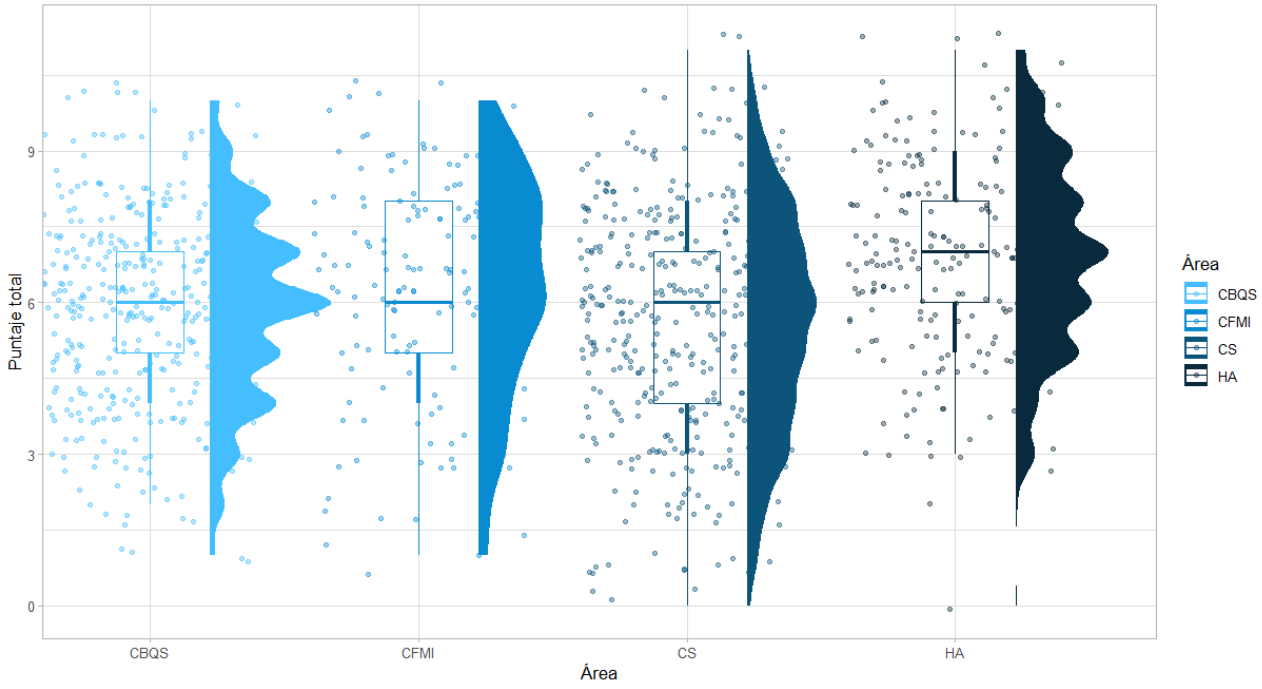
1. Examen de Comprensión Lectora

1.1 Análisis descriptivo

Para el análisis descriptivo, se comenzó por generar la Figura 1 en la que se puede apreciar la distribución en el puntaje total para cada una de las áreas del posgrado. Se observan distribuciones parecidas en las cuatro áreas, aunque se destaca que sí se observan diferencias en la tendencia central de cada área, siendo el área HA la que presenta puntajes más altos en términos generales. Estos resultados fueron consistentes en una prueba ANOVA que resultó significativa (F = 20.23; p< .05). Al analizarse mediante una prueba de Tukey, se encontró que las diferencias significativas correspondían a la comparación entre el área HA y el resto de las áreas.

Figura 1.

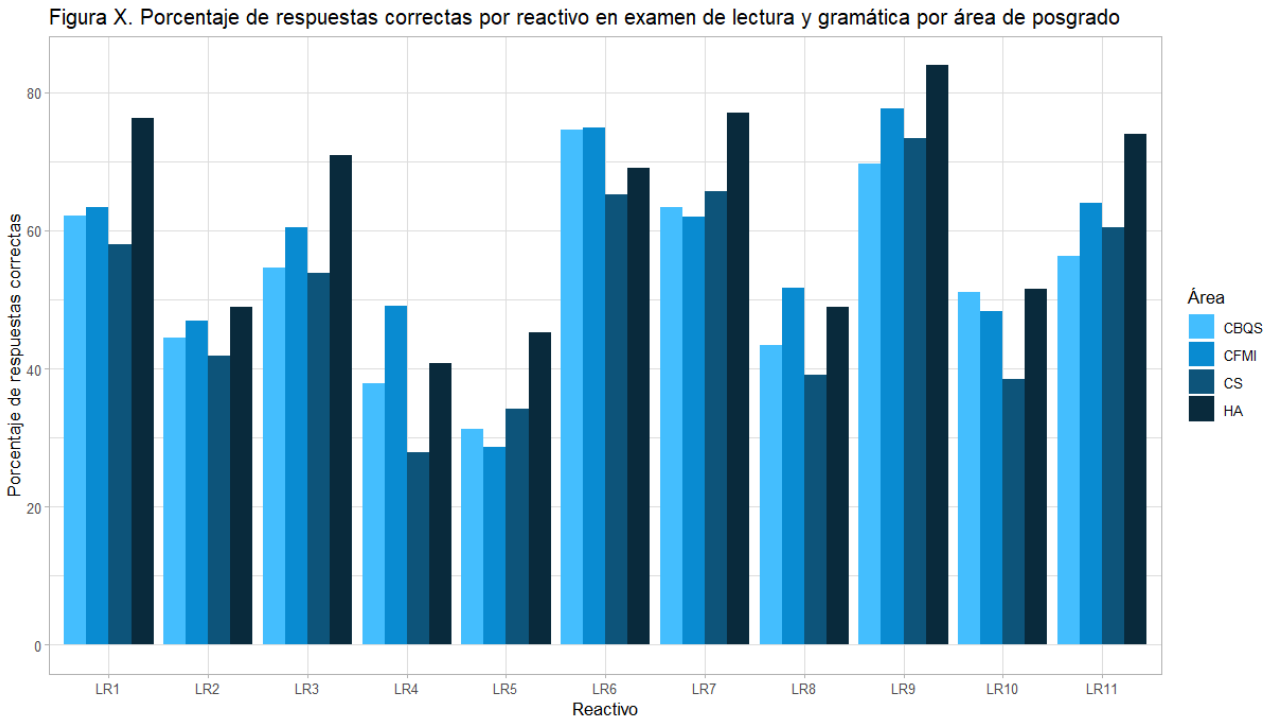
Distribución de puntajes en examen de comprensión lectora por área de posgrado



De igual forma, en la Figura 2, se analizó de manera directa el porcentaje de respuestas correctas en cada uno de los reactivos del examen por cada área del posgrado. Nuevamente se observa que, en términos generales, el área HA es la que presenta mejores resultados, seguida del área CFMI.

Figura 2.

Porcentaje de respuestas correctas por reactivo en examen de comprensión lectora por área de posgrado



1.2 Análisis DIF con métodos no paramétricos

1.2.1 Método de Mantel-Haenszel

El análisis DIF mediante el método de Mantel-Haenszel se realizó primero por pares de áreas de conocimiento del posgrado. En la Tabla 3 se puede observar de manera sintética los resultados del mismo. Se presentan en cada columna los resultados de cada análisis por pares, mostrando el valor de deltaMH y su interpretación mediante códigos de signos positivos y negativos. Los valores positivos indican que el reactivo resulta más sencillo al primero de los dos grupos de cada comparación, mientras que los negativos indican que resulta más sencillo para el

segundo grupo. 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 3.

Análisis DIF con método Mantel-Haenszel área contra área en examen de comprensión lectora

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
LR1	0.353	0.242	1.088 +	0.007	0.675	0.725
LR2	0.113	-0.091	-0.102	-0.226	-0.326	-0.103
LR3	-0.313	-0.023	0.689	0.217	0.986	0.664
LR4	-0.634	-1.663 --	-1.658 --	-1.063 -	-0.938	-0.085
LR5	0.586	1.056 +	1.409 +	0.467	0.694	0.412
LR6	0.232	-0.411	-1.378 -	-0.668	-1.790 --	-1.107 -
LR7	0.260	1.053 +	1.188 +	0.675	0.855	0.093
LR8	-0.360	-0.637	-0.954	-0.106	-0.544	-0.203
LR9	-0.876	0.369	-0.022	1.119 +	0.949	-0.420
LR10	1.156 +	-0.275	-0.229	-1.254 -	-1.259 -	0.080
LR11	-0.406	0.577	0.454	0.824	0.927	-0.064

Dentro de este análisis, 7 de los 11 reactivos del examen de comprensión lectora presentan DIF en al menos una de las combinaciones comparativas entre las áreas (reactivos 1, 4, 5, 6, 7, 9, 10). Se puede observar que en la comparativa directa, los reactivos con mayor presencia de DIF significativo son los reactivos 4, 5 y 6, siendo el reactivo 4 el que presenta mayor nivel de DIF, especialmente al comparar el área FMI con las áreas CS y HA. Cabe destacar que los reactivos 2, 3, 8 y 11 no presentan DIF significativo en ninguna de las comparativas. Por otro lado, las comparativas entre FMI y CBQS y CS con HA son las que menos reactivos con DIF presentan.

1.2.2 Método de regresión logística

Como se observa en la Tabla 4, en el análisis DIF mediante el método de regresión logística, los reactivos 4, 5 6, y 7 son los que presentan un mayor nivel de DIF uniforme, aunque también se encontró presencia importante de DIF en los reactivos 9 y 11. Se destaca nuevamente que la mayor presencia de DIF surge al comparar las áreas FMI y CBQS con las áreas CS y HA.

Tabla 4.

Análisis DIF con método de Regresión Logística para DIF uniforme área contra área en examen de comprensión lectora

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
LR1	0.001	0.001	0.010	0.000	0.004	0.003
LR2	0.000	0.000	0.000	0.000	0.001	0.000
LR3	0.001	0.000	0.004	0.000	0.010 *	0.003
LR4	0.005	0.020 ***	0.025 **	0.010 **	0.007 *	0.000
LR5	0.004	0.010 *	0.024 **	0.003	0.006	0.002
LR6	0.000	0.001	0.018 *	0.004	0.028 ***	0.010 *
LR7	0.002	0.011 *	0.018 *	0.004	0.008 *	0.000
LR8	0.001	0.003	0.009	0.000	0.002	0.000
LR9	0.004	0.001	0.000	0.010 **	0.009 *	0.000
LR10	0.007 *	0.000	0.001	0.014 ***	0.014 **	0.000
LR11	0.002	0.002	0.002	0.008 *	0.010 *	0.000

En cuanto a la presencia de DIF no uniforme con el método de regresión logística presentado en la Tabla 5, se encontró que este tipo de DIF es menos común dentro de este examen, apareciendo únicamente en los reactivos 5, 6 y 11. Cabe destacar que en este caso, el patrón es distinto a lo visto con el DIF uniforme. En particular, dos de las tres ocurrencias de DIF se dieron en la comparación entre las áreas CBQS y FMI y las áreas CS y HA.

Tabla 5.

Análisis DIF con método de Regresión Logística para DIF no uniforme área contra área en examen de comprensión lectora

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
LR1	0.003	0.002	0.009	0.000	0.001	0.002
LR2	0.003	0.002	0.008	0.000	0.001	0.001
LR3	0.002	0.000	0.001	0.005	0.005	0.000
LR4	0.000	0.005	0.004	0.005	0.003	0.000
LR5	0.001	0.001	0.008	0.003	0.003	0.013 **
LR6	0.002	0.002	0.004	0.000	0.011 *	0.012 *
LR7	0.001	0.000	0.001	0.002	0.002	0.000
LR8	0.000	0.004	0.001	0.002	0.000	0.002
LR9	0.005	0.000	0.004	0.005	0.000	0.002
LR10	0.001	0.003	0.002	0.000	0.000	0.000
LR11	0.010 *	0.003	0.006	0.003	0.001	0.000

1.2.3 Modelo Exploratorio de Regresión Logística

El análisis DIF con el método de regresión logística exploratorio permitió analizar si la distinción entre las distintas áreas de posgrado era la única fuente relevante de DIF, y de ser así, entre qué áreas se encontraría la mayor diferencia. Para este análisis, se incluyeron como covariables el área del posgrado, el nivel del posgrado (maestría o doctorado), los posgrados en específico y el género de los sustentantes. Con estas covariables, como se observa en la Tabla 6, se detectó presencia de DIF uniforme en los reactivos 4 y 11 dividiendo por un lado a las áreas 1 y 2 y por otro a las áreas 3 y 4. Además de esto, se encontró presencia de DIF no uniforme en los reactivos 6 y 10 con la misma división entre áreas. Finalmente se detectó presencia de DIF uniforme en el reactivo 5 por el posgrado específico, aunque el particionamiento acabó agrupando como un conjunto a los posgrados referentes a áreas CFMI y CBQS y otro a los posgrados referentes a las áreas CS y HA.

Tabla 6.

Análisis DIF con método exploratorio de Regresión Logística para DIF en examen de comprensión lectora

Reactivo	DIF	Tipo	Variables	Particiones
LR1	No	---	---	---
LR2	No	---	---	---
LR3	No	---	---	---
LR4	Sí	Uniforme	Área	1
LR5	Sí	Uniforme	Posgrado	1
LR6	Sí	Uniforme	Área	1
LR7	No	---	---	---
LR8	No	---	---	---
LR9	No	---	---	---
LR10	Sí	Uniforme	Área	1
LR11	Sí	Uniforme	Posgrado	1

1.3 Análisis DIF con métodos IRT basados en modelo de Rasch

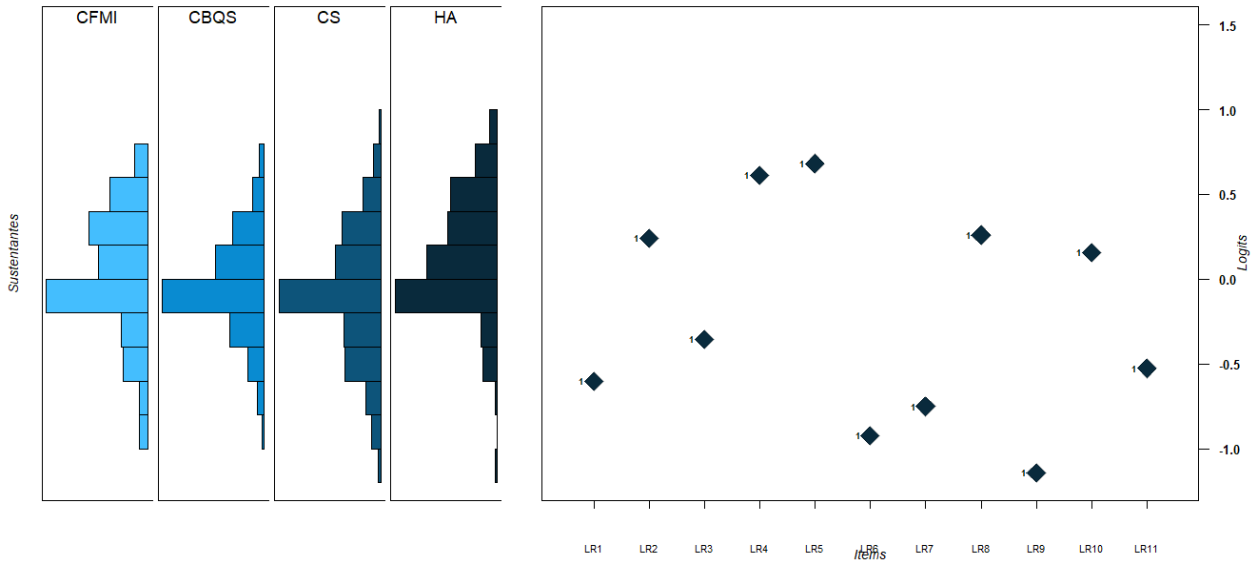
1.3.1 Mapa de Wright

Para visualizar la distribución de la dificultad de los reactivos a partir de un modelo de Rasch, así como la distribución de los participantes, se presenta en la Figura 3 el mapa de Wright por área de conocimiento. Se observa que en general, la distribución de los sustentantes es muy

similar en las cuatro áreas, aunque se observa una ligera agrupación en niveles de habilidad más altos en el Área HA.

Figura 3.

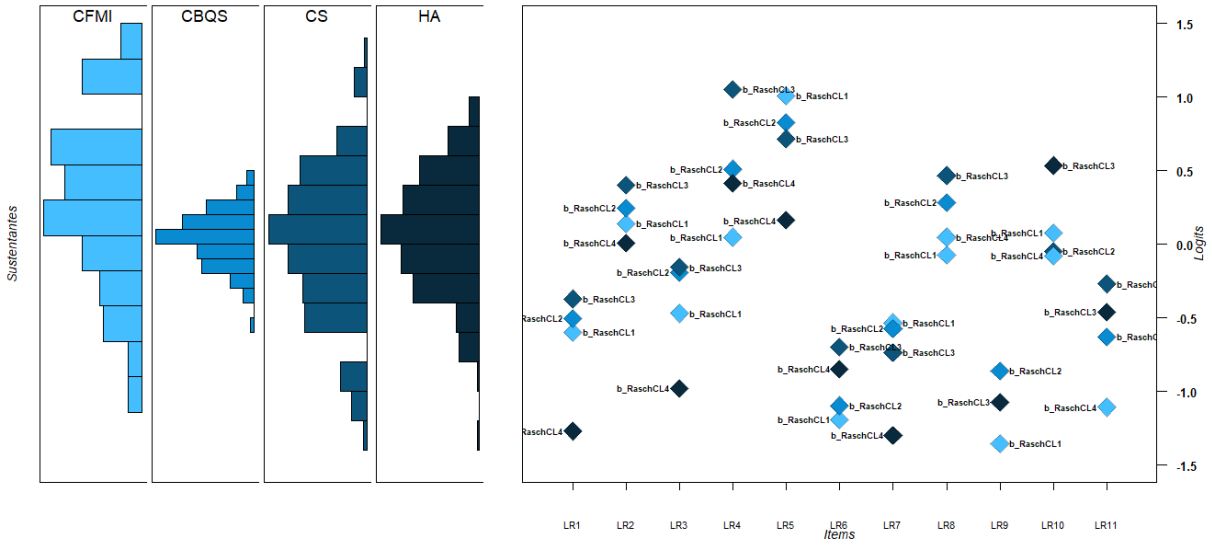
Mapa de Wright del examen de Comprensión Lectora con distribución por área



Para poder contrastar también si había diferencias en la dificultad de los reactivos, se calculó el modelo de Rasch permitiendo que los parámetros de dificultad variasen entre cada área de conocimiento. Los resultados se pueden observar en la Figura 4. Se observan diferencias importantes en la dificultad de reactivos como el 1, 5, 7 y 11. Además se observa que en general, el parámetro de dificultad suele ser más bajo para el Área HA.

Figura 4.

Mapa de Wright del examen de Comprensión Lectora con dificultad calculada por área

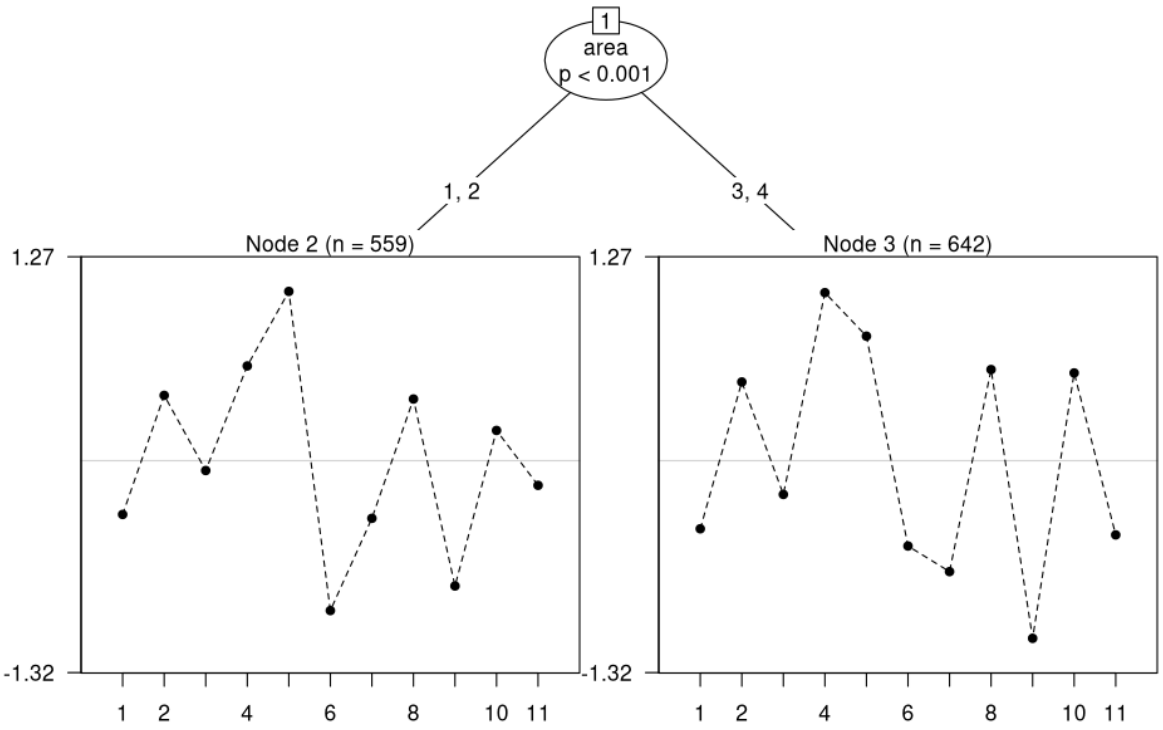


1.3.2 Análisis DIF exploratorio

Para evaluar nuevamente si la presencia de DIF era específica de las diferencias en las áreas de posgrado, pero a partir de un modelo paramétrico, se utilizó el método Raschtree del paquete *psychotree.* Las covariables incluidas fueron el género, el área del posgrado y el posgrado específico al que aplicaban los sustentantes. Los resultados se pueden observar en la Figura 5. Como se observa, este método exploratorio particionó los datos únicamente a partir del área de posgrado, encontrando presencia de DIF en la comparación de las áreas CFMI y CBQS contra las áreas CS y HA.

Figura 5.

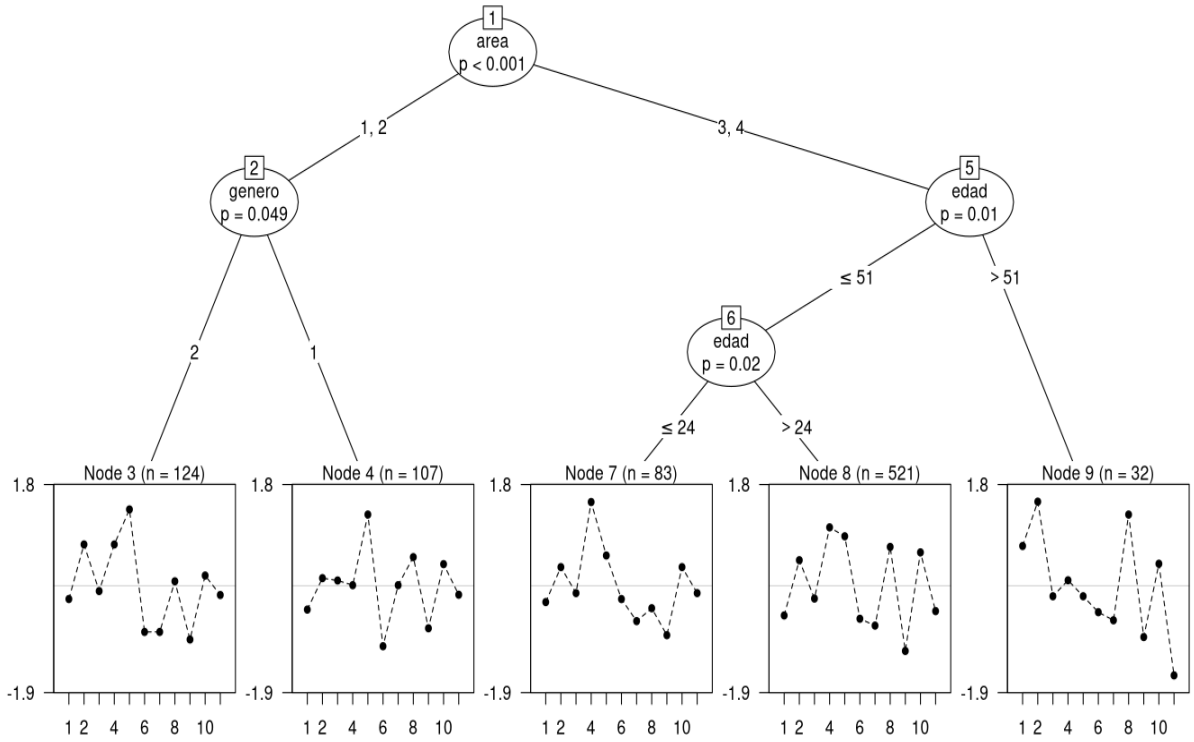
Particionamiento de estimación de parámetros de dificultad por área con modelo de Rasch y método Raschtree



Cabe destacar que, al incluir la variable de la edad, se observaron ramificaciones distintas, aunque en ambos casos, se parte de la diferencia entre las áreas de posgrado CFMI y CBQS con las áreas CS y HA. Se puede observar la diferencia observada con la variable edad en la Figura 6.

Figura 6.

Particionamiento de estimación de parámetros de dificultad por área, género y edad con modelo de Rasch y método Raschtree



1.3.4 Método de diferencia de logits con modelo de Rasch

En cuanto al análisis DIF específico a partir de estos modelos, se puede observar en la Tabla 6 que los reactivos con mayor presencia de DIF son los reactivos 4, 5 y 6, y las mayores diferencias se encuentra en la comparación entre FMI y CS y la comparación entre FMI y HA.

Tabla 7.

Análisis DIF exploratorio del examen de lectura con método de Rasch

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
LR1	0.083	0.071	0.390	-0.012	0.307	0.319
LR2	0.069	0.036	-0.154	-0.033	-0.223	-0.190
LR3	-0.101	-0.014	0.228	0.087	0.329	0.242
LR4	-0.286	-0.699 -	-0.648 -	-0.413	-0.363	0.051
LR5	0.342	0.571 +	0.544 +	0.230	0.203	-0.027
LR6	0.071	-0.206	-0.639 -	-0.277	-0.710 -	-0.433
LR7	0.216	0.503 +	0.481	0.287	0.265	-0.021
LR8	-0.180	-0.241	-0.401	-0.061	-0.221	-0.160
LR9	-0.330	0.009	0.121	0.339	0.452	0.113
LR10	0.303	-0.159	-0.120	-0.462	-0.424	0.038
LR11	-0.186	0.129	0.198	0.315	0.384	0.069

1.3.5 Análisis DIF con método de χ^2 de Lord

Finalmente, dentro de los métodos convencionales para detección de DIF paramétricos, se utilizó el método de χ^2 de Lord. Los resultados se pueden observar en la Tabla 7. Bajo este método, se observa una mayor presencia de DIF, teniendo hasta 8 reactivos señalados en tres o más comparaciones (reactivos 1, 4, 5, 6, 8, 9, 10 y 11). Se observa que las comparaciones donde hubo mayor presencia de DIF son las referentes a las áreas CBQS y CS y FMI y CBQS.

Tabla 8.

Análisis DIF exploratorio del examen de lectura con método χ^2de Lord

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
LR1	43.669 ***	0.315	8.655 *	82.15 ***	1.941	9.078 *
LR2	0.968	2.165	0.979	3.096	0.436	1.882
LR3	4.959	1.355	8.187 *	11.484 **	4.725	5.614
LR4	60.287 ***	13.872 ***	78.841 ***	33.746 ***	6.737 *	41.353 ***
LR5	16.412 ***	0.474	1.047	328.821 ***	11.762 **	2.068
LR6	65.079 ***	1.436	5.06	96.651 ***	4.272	16.598 ***
LR7	10.543 **	1.57	0.495	34.663 ***	0.564	2.073
LR8	127.019 ***	8.85 *	32.649 ***	9.631 **	5.696	5.919
LR9	92.59 ***	0.516	3.907	194.883 ***	3.605	10.333 **
LR10	104.061 ***	6.093 *	52.907 ***	26.175 ***	8.992 *	21.821 ***
LR11	131.557 ***	3.837	5.955	115.364 ***	2.519	7.607 *

1.3.6 Comparación de 2 grupos

Dados los resultados de los métodos exploratorios y las hipótesis generadas a priori, se decidió comparar de forma directa las áreas CFMI y CBQS contra las áreas CS y HA, generando así un análisis de 2 grupos a partir de los modelos ya presentados anteriormente. Como se puede observar en la Tabla 8, los reactivos 4, 6 y 11 son señalados con presencia de DIF en 3 de los cuatro métodos utilizados. Cabe destacarse que, bajo esta propuesta, con el método de diferencia de logits de Rasch, ningún reactivo fue señalado con presencia de DIF.

Tabla 9.

Comparación de dos grupos con métodos tradicionales

Reactivo	Mantel-Haenszel	Regresión Logistica	Diferencia de Logits (Rasch)	χ^2 de Lord
LR1	-0.313	0.001	0.106	6.772 *
LR2	0.191	0.000	-0.083	4.151
LR3	-0.358	0.001	0.142	1.396
LR4	1.221 +	0.010 ***	-0.465	75.544 ***
LR5	-0.773	0.006 **	0.309	0.797
LR6	1.017 +	0.010 **	-0.400	12.145 **
LR7	-0.824	0.005 **	0.335	3.339
LR8	0.356	0.001	-0.163	29.306 ***
LR9	-0.853	0.006 *	0.297	15.699 ***
LR10	1.000 +	0.007 **	-0.371	58.584 ***
LR11	-0.755	0.005 **	0.293	10.746 **

1.4 Métodos paramétricos robustos - IRT

1.4.1 Selección de Modelo

Para el análisis con métodos paramétricos más robustos, se comenzó por determinar qué modelo era el más idóneo para representar los datos. Como se observa en la Tabla 9, de los modelos comparados, únicamente el modelo de Rasch presenta un ajuste inadecuado, al tener un estadístico M2 con una $p<.05$ y con valores para CFI y RMSEA mayores a los puntos de corte normalmente aceptables. Por otro lado, cabe destacar que no parece haber una diferencia realmente importante entre el modelo logístico de 2 parámetros y el de 3 parámetros, bajo el entendido de que el añadir parámetros al modelo genera un mejor ajuste invariablemente, la mejora no parece ser sustancial.

Un análisis a partir del test de Razón de Verosimilitud confirma que el modelo 2PL resulta ser más adecuado.

Tabla 10.

Comparación de modelos TRI unidimensionales

Modelos	M2	df	*p*	RMSEA	CFI
Modelo Rasch	131.656	54	0.000	0.035	0.825
Modelo 2PL	53.053	44	0.165	0.013	0.980
Modelo 3PL	40.699	33	0.168	0.014	0.983

Posteriormente, una vez determinado el mejor modelo a utilizar, se procedió a valorar el supuesto de unidimensionalidad para determinar si era necesario utilizar modelos multidimensionales para este examen. Para ello, se comenzó por comparar de manera directa el modelo de 2PL unidimensional previamente descrito con un modelo de 2 parámetros con 2 dimensiones. Como se observa en la Tabla 10, el test de razón de verosimilitud resultó significativo por un margen mínimo bajo el criterio de $p<.05$. Además de eso, los valores de los estadísticos de BIC y AIC apuntan a que resulta mejor el modelo unidimensional. A partir de esto, se asume el cumplimiento del supuesto de unidimensionalidad y se procede a utilizar el modelo unidimensional logístico de 2 parámetros.

Tabla 11.

Comparación de modelos TRI unidimensionales mediante Test de Razón de Verosimilitud

Modelos	AIC	BIC	logLik	X2	df	*p*
Modelo Unidimensional	17,176.40	17,288.71	−8,566.20			
Modelo exploratorio de dos dimensiones	17,177.41	17,340.77	−8,556.71	18.99	10	0.04036

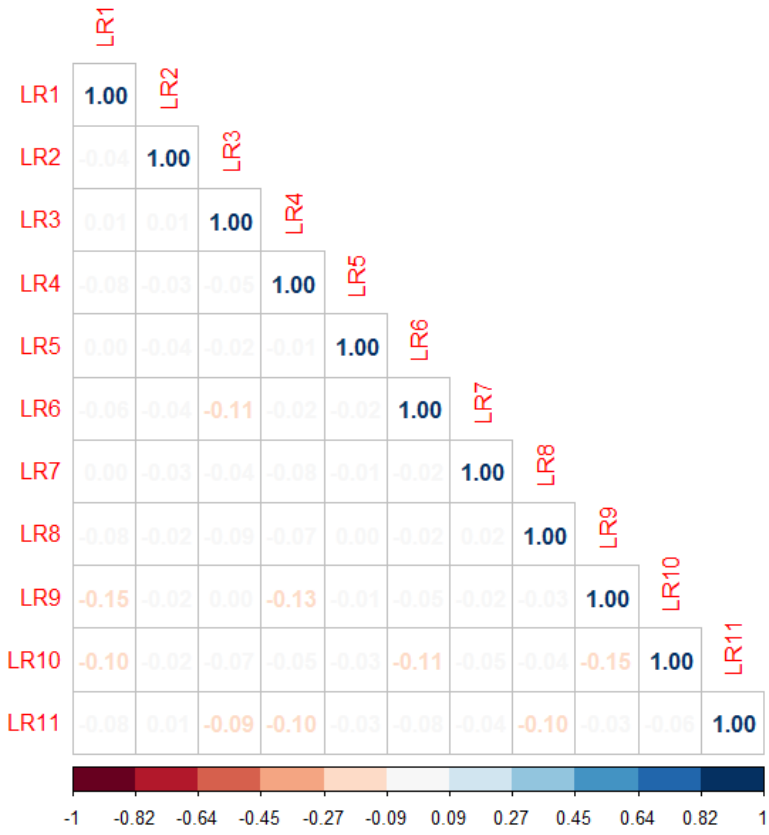
1.4.2 Verificación de supuestos

Dado que el nivel de ajuste del modelo unidimensional de 2PL resulta adecuado, es pertinente asumir que tanto el supuesto de monotonicidad como el de unidimensionalidad se cumplen. Por ello, únicamente resta verificar en este modelo el supuesto de independencia local.

Para ello, se utilizó el estadístico Q3, que permite comparar las correlaciones entre los residuales de los reactivos. Estas correlaciones pueden ser observadas de forma gráfica en la Figura 7.

Figura 7.

Matriz de valores residuales del test Q3 para evaluar independencia local en examen de Comprensión Lectora



Como se puede observar, ninguna de las correlaciones resulta mayor a .2, por lo que estos resultados apoyan la evidencia respecto al mantenimiento del supuesto de independencia local.

1.4.3 Método de razón de verosimilitud

Para el método DIF-Free-Then-DIF, se llevó a cabo un primer análisis con el test de Razón de Verosimilitud, encontrando como único reactivo libre de DIF el reactivo 2; aunque el 8 presentó un valor de $p = .04$, se procedió a utilizar únicamente el reactivo 2 como reactivo ancla.

En la Tabla 11, se observan los resultados obtenidos de este análisis. Dado que se trata de un test de Razón de Verosimilitud que compara directamente el nivel de ajuste de un modelo donde los parámetros son iguales para las cuatro áreas y otro donde los parámetros pueden variar entre éstas, se espera que cualquier valor de p<-05 indique la presencia de DIF. Cabe destacar que, pese a que no se encontraron reactivos con DIF significativo, la mayoría no convergió dentro del análisis iterativo, esto puede ser debido al tamaño de muestra, pues la complejidad de este tipo de análisis suele requerir muestras más grandes. Estudios como el de Rogers y Swaminathan (1993) han mostrado que, incluso para análisis más simples como el de regresión logística, 250 participantes por grupo generan dificultades y errores de medición importantes, recomendando un mínimo de 500 participantes por grupo.

Tabla 12.

Análisis DIF con método de Razón de Verosimilitud en examen de comprensión lectora

Reactivo	Convergencia	AIC	SABIC	HQ	BIC	X2	*p*
LR1		11.257	22.828	22.786	41.886	0.743	0.994
LR3		10.115	21.687	21.645	40.745	1.885	0.93
LR4		10.089	21.660	21.618	40.719	1.911	0.928
LR5		3.407	14.978	14.937	34.037	8.593	0.198
LR6		5.857	17.428	17.386	36.487	6.143	0.407
LR7		8.226	19.797	19.755	38.856	3.774	0.707
LR8	Convergió	8.079	19.651	19.609	38.709	3.921	0.687
LR9	Convergió	8.868	20.440	20.398	39.498	3.132	0.792
LR10	Convergió	9.116	20.687	20.645	39.746	2.884	0.823
LR11		7.903	19.474	19.432	38.532	4.097	0.664

Dado que el análisis de interés es en el contraste directo área por área, en la Tabla 12 se presentan los resultados de este mismo análisis, pero en el contraste por pares de áreas. El valor presentado es el del estadístico X2. Dado que nuevamente para este caso se siguió la estrategia DIF-Free-Then-DIF, se presentan únicamente los resultados de los reactivos que no se seleccionaron como reactivos ancla en un primer análisis.

Tabla 13.

Análisis DIF con método de Razón de Verosimilitud área contra área en examen de comprensión lectora

Reactivo	FMI vs CBQS	FMI vs CS	FMI vs HA	CBQS vs CS	CBQS vs HA	CS vs HA	Todos los grupos
CL1	-	-	8.94 *	-	11.19 **	0.33	0.74
CL2	-	-	-	-	-	-	
CL3	-	-	-	2.45	17.92 ***	0.55	1.88
CL4	2.62	11.27 **	-	18.8 ***	-	0.57	1.91
CL5	-	-	15.16 ***	-	14.8 ***	4.3	8.59
CL6	-	-		6.82 *	-	7.6 *	6.14
CL7	-	-	9.38 **	-	12.33 **	0.66	3.77
CL8	-	4.86	-	-	-	-	**3.92**
CL9	-	-	-	-	12.59 **	1.4	**3.13**
CL10	-	-	-	17.88 ***	-	0.15	**2.88**
CL11	2.848	-	3.22	-	15.38 ***	0.6	4.09

Como se puede observar en la Tabla 12, los reactivos con mayor nivel de DIF son el reactivo 4, 5 y el 7. Además, se destaca nuevamente, que la comparación con mayor presencia de DIF es entre el área CBQS y el área HA, de manera consistente con los análisis realizados previamente.

1.4.4 Análisis DIF con dos grupos

Finalmente, se presenta en la Tabla 13 el análisis correspondiente a comparar las áreas CFMI y CBQS agrupadas contra las áreas CS y HA. En el primer análisis con todos los reactivos, se encontró únicamente ausencia de DIF significativo en los reactivos 1, 2 y 8. Posteriormente, en un segundo análisis utilizando estos tres reactivos como reactivos ancla, del resto, los reactivos 3, 9, 10 y 11 presentaron una $p>.05$, por lo que se descartó la presencia de DIF significativo en estos reactivos, señalando únicamente a los reactivos 4, 5, 6 y 7 con presencia de DIF significativo.

Tabla 14.

Análisis DIF con método de Razón de Verosimilitud en examen de comprensión lectora - Comparación con 2 grupos

Reactivos	Convergencia	AIC	SABIC	HQ	BIC	X2	*p*
LR3	Convergió	0.232	4.089	4.075	10.442	3.768	0.152
LR4	Convergió	−9.411	−5.554	−5.568	0.799	13.411	0.001
LR5	Convergió	−5.328	−1.471	−1.485	4.882	9.328	0.009
LR6	Convergió	−3.198	0.659	0.645	7.012	7.198	0.027
LR7	Convergió	−3.228	0.629	0.615	6.982	7.228	0.027
LR9	Convergió	−0.713	3.144	3.130	9.497	4.713	0.095
LR10	Convergió	−1.461	2.396	2.382	8.749	5.461	0.065
LR11	Convergió	0.429	4.287	4.273	10.639	3.571	0.168

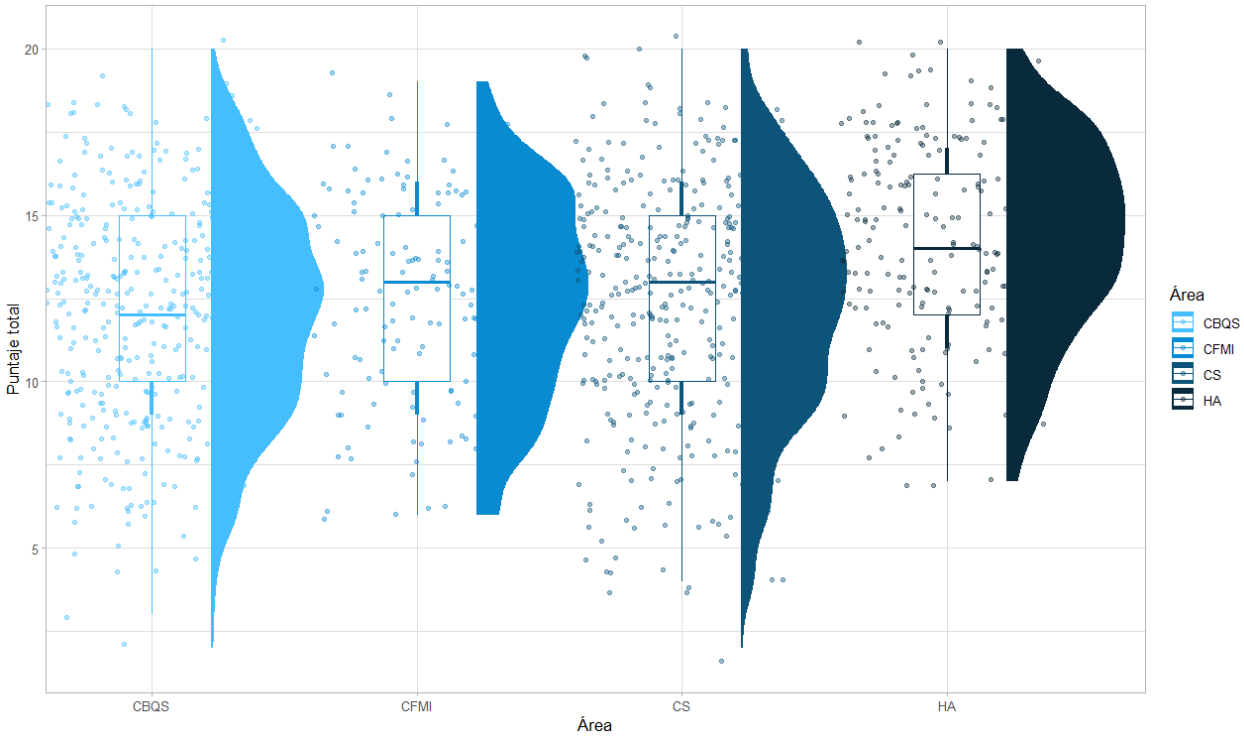
2. Examen de Redacción y Gramática

2.1 Análisis descriptivo

Como se muestra en la Figura 8, las distribuciones en el puntaje total para cada una de las áreas del posgrado son similares en las cuatro áreas, aunque al igual que en el examen de comprensión lectora, se observa una tendencia de puntajes más altos en el área HA, seguida por el área CFMI.

Figura 8.

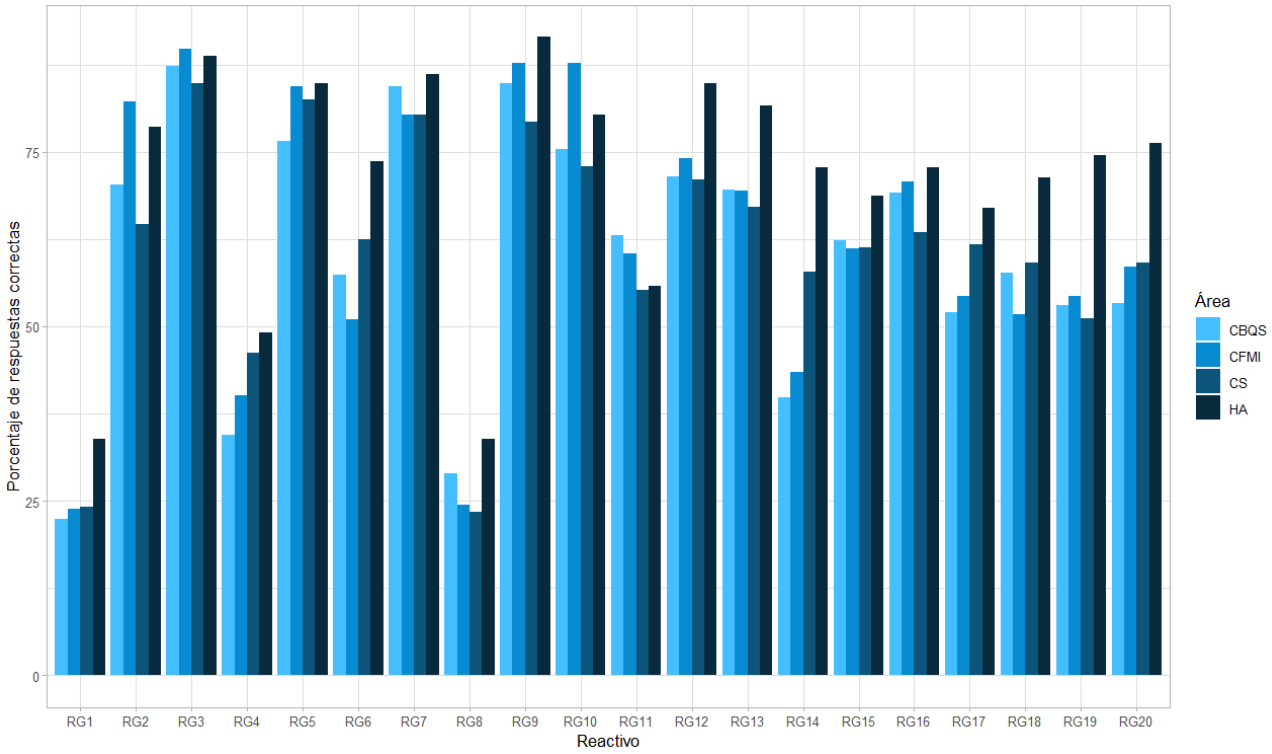
Distribución de puntajes en examen de Redacción y Gramática por área de posgrado



De igual forma, en la Figura 9, se analizó de manera directa el porcentaje de respuestas correctas en cada uno de los reactivos del examen por cada área del posgrado. Nuevamente se observa que, en términos generales, el área HA es la que presenta mejores resultados, seguida del área CFMI.

Figura 9.

Porcentaje de respuestas correctas por reactivo en examen de redacción y gramática por área de posgrado



2.2 Análisis DIF con métodos no paramétricos

2.2.1 Método de Mantel-Haenszel

En este análisis presentado en la Tabla 14, se puede observar la presencia de DIF en al menos una de las comparativas en 15 de los 20 reactivos (2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 17, 18, 19 y 20). En este caso, se observa que los reactivos con mayor presencia de DIF a través de distintas comparaciones son los reactivo 6, 10, y 14, siendo los únicos que presentan DIF en 3 o más de las comparativas. De igual forma, tal como sucedió en el examen de comprensión lectora, se observa que las comparativas con menor presencia de reactivos con DIF son la comparación entre FMI y CBQS y entre CS y HA.

Tabla 15.

Análisis DIF con Método de Mantel-Haenszel en examen de Redacción y Gramática

Reactivo	FMI y CBQS		FMI y CS		FMI y HA		CBQS y CS		CBQS y HA		CS y HA	
RG1	-0.113		0.187		0.028		0.176		0.189		-0.200	
RG2	-1.679	--	-2.219	--	-1.923	--	-0.791		-0.585		0.571	
RG3	-0.153		-0.926		-1.940	--	-0.587		-1.426	-	-0.712	
RG4	-0.406		0.638		-0.176		1.177	+	0.272		-0.788	
RG5	-1.068	-	-0.092		-1.617	--	0.987		-0.349		-0.979	
RG6	1.049	+	1.632	++	1.703	++	0.446		0.497		0.029	
RG7	1.187	+	0.296		-0.316		-0.739		-1.308	-	-0.228	
RG8	0.745		0.065		0.058		-0.895		-0.572		0.089	
RG9	-0.313		-1.451	-	-0.448		-1.099	-	0.205		1.247	+
RG10	-2.096	--	-2.278	--	-2.462	--	-0.405		-0.547		0.028	
RG11	0.348		-0.485		-1.040	-	-0.918		-1.593	--	-0.853	
RG12	0.035		-0.358		0.039		-0.170		0.324		0.595	
RG13	0.152		-0.078		1.035	+	-0.308		0.332		0.746	
RG14	-0.183		1.666	++	2.153	++	1.963	++	2.444	++	0.456	
RG15	0.172		0.159		-0.151		-0.119		-0.357		-0.178	
RG16	-0.133		-0.702		-0.818		-0.669		-0.807		0.037	
RG17	0.065		1.107	+	0.156		1.040	+	0.295		-0.929	
RG18	0.900		0.954		1.063	+	0.056		0.016		0.039	
RG19	0.351		-0.159		1.018	+	-0.495		0.715		1.140	+
RG20	-0.129		0.293		0.523		0.542		1.069	+	0.455	

2.2.2 Método de regresión logística

Al evaluar la presencia de DIF uniforme en la Tabla 15, los resultados son sumamente similares a los identificados en el método MH, observando mayor presencia de DIF en los reactivos 2, 6, 10 y 14. Cabe destacar que, si bien nuevamente las comparaciones con más cantidad de reactivos con DIF son la comparación entre FMI y CS y FMI y HA, también se observa presencia importante de DIF en la comparación entre CBQS y CS.

Tabla 16.

Análisis DIF exploratorio del examen de redacción y gramática con método de regresión logística para DIF uniforme

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
RG1	0.000	0.000	0.000	0.000	0.001	0.000
RG2	0.014 *	0.035 ***	0.031 **	0.007 *	0.002	0.003
RG3	0.000	0.003	0.024 *	0.002	0.012 *	0.004
RG4	0.002	0.005	0.000	0.015 ***	0.001	0.006
RG5	0.006	0.000	0.017 *	0.009 *	0.001	0.008
RG6	0.008 *	0.018 **	0.024 **	0.002	0.002	0.000
RG7	0.008	0.001	0.000	0.004	0.010 *	0.000
RG8	0.005	0.000	0.000	0.008 *	0.002	0.000
RG9	0.000	0.010 *	0.001	0.009 *	0.000	0.010 *
RG10	0.024 **	0.035 ***	0.048 ***	0.002	0.002	0.000
RG11	0.001	0.002	0.016 *	0.010 *	0.026 ***	0.008 *
RG12	0.000	0.000	0.000	0.000	0.001	0.002
RG13	0.000	0.000	0.005	0.002	0.001	0.004
RG14	0.000	0.025 ***	0.048 ***	0.040 ***	0.048 ***	0.002
RG15	0.001	0.000	0.000	0.000	0.001	0.000
RG16	0.000	0.004	0.004	0.006 *	0.006	0.000
RG17	0.000	0.010 *	0.001	0.012 **	0.001	0.007 *
RG18	0.007	0.008	0.008	0.000	0.000	0.000
RG19	0.000	0.000	0.006	0.001	0.003	0.007 *
RG20	0.001	0.001	0.003	0.004	0.008 *	0.002

Nuevamente, como se observó en el examen de comprensión lectora, en su mayoría la presencia de DIF es de carácter uniforme, pues solamente tres de los 20 reactivos aquí analizados presentaron DIF no uniforme. De manera destacable, se observa que el reactivo 1 presenta DIF no uniforme en múltiples comparaciones (Tabla 16).

Tabla 17.

Análisis DIF exploratorio del examen de lectura con método de regresión logística para DIF no uniforme

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
RG1	0.000	0.004	0.027 **	0.007 *	0.029 ***	0.009 *
RG2	0.000	0.002	0.000	0.005	0.000	0.004
RG3	0.004	0.006	0.005	0.000	0.000	0.000
RG4	0.001	0.000	0.000	0.001	0.000	0.000
RG5	0.005	0.009	0.004	0.001	0.000	0.002
RG6	0.001	0.001	0.000	0.000	0.000	0.001
RG7	0.000	0.004	0.003	0.003	0.001	0.000
RG8	0.000	0.005	0.003	0.008 *	0.004	0.001
RG9	0.006	0.008	0.014	0.000	0.001	0.000
RG10	0.001	0.000	0.006	0.001	0.002	0.006
RG11	0.001	0.006	0.010	0.002	0.004	0.000
RG12	0.003	0.005	0.002	0.000	0.000	0.001
RG13	0.005	0.005	0.008	0.000	0.000	0.000
RG14	0.000	0.001	0.006	0.000	0.003	0.002
RG15	0.001	0.001	0.000	0.000	0.001	0.001
RG16	0.001	0.000	0.000	0.001	0.001	0.000
RG17	0.000	0.000	0.000	0.001	0.000	0.001
RG18	0.000	0.000	0.008	0.000	0.008 *	0.012 **
RG19	0.000	0.000	0.001	0.000	0.000	0.000
RG20	0.004	0.006	0.002	0.000	0.001	0.002

2.2.3 Modelo Exploratorio de Regresión Logística

El análisis DIF con el método de regresión logística y con las mismas covariables detectó presencia de DIF uniforme en el reactivo 4 entre las áreas CFMI y CBQS y las áreas CS y HA. Además de esto, encontró presencia de DIF no uniforme en los reactivos 6 y 10 con la misma división entre áreas. Finalmente encontró presencia de DIF uniforme por el posgrado específico, aunque el particionamiento mostró como un grupo a los posgrados pertenecientes a áreas CFMI y CBQS y como otro grupo a los posgrados de las áreas CS y HA.

Tabla 18.

Análisis DIF con método exploratorio de Regresión Logística para DIF en examen de redacción y gramática

Reactivo	DIF	Tipo	Variables	Particiones
RG1	Sí	No-uniforme	Área	1
RG2	No	---	---	---
RG3	No	---	---	---
RG4	No	---	---	---
RG5	No	---	---	---
RG6	No	---	---	---
RG7	No	---	---	---
RG8	No	---	---	---
RG9	No	---	---	---
RG10	No	---	---	---
RG11	No	---	---	---
RG12	No	---	---	---
RG13	No	---	---	---
RG14	No	---	---	---
RG15	No	---	---	---
RG16	No	---	---	---
RG17	No	---	---	---
RG18	No	---	---	---
RG19	No	---	---	---
RG20	No	---	---	---

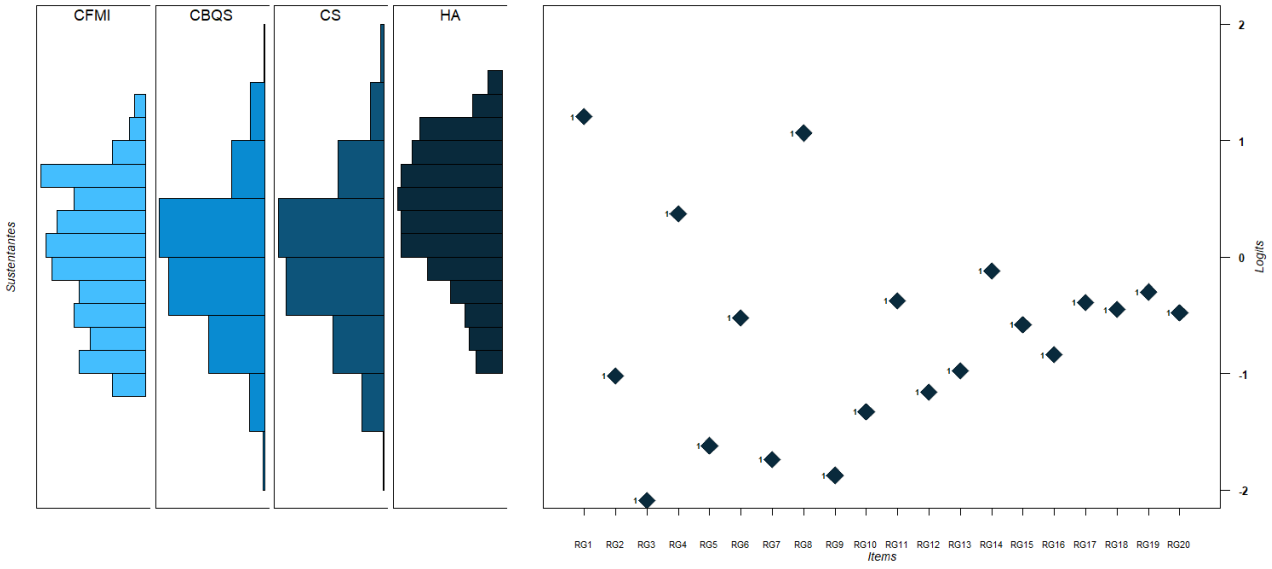
2.3 Análisis DIF con métodos IRT basados en modelo de Rasch

2.3.1 Mapa de Wright

Para visualizar de mejora manera la distribución de la dificultad de los reactivos, así como la distribución de los participantes, se presenta en la Figura 10 el mapa de Wright por área de conocimiento. Se observa que en general, la distribución de los sustentantes es muy similar en las cuatro áreas, aunque se observa una ligera agrupación en niveles de habilidad más altos en el Área HA.

Figura 10.

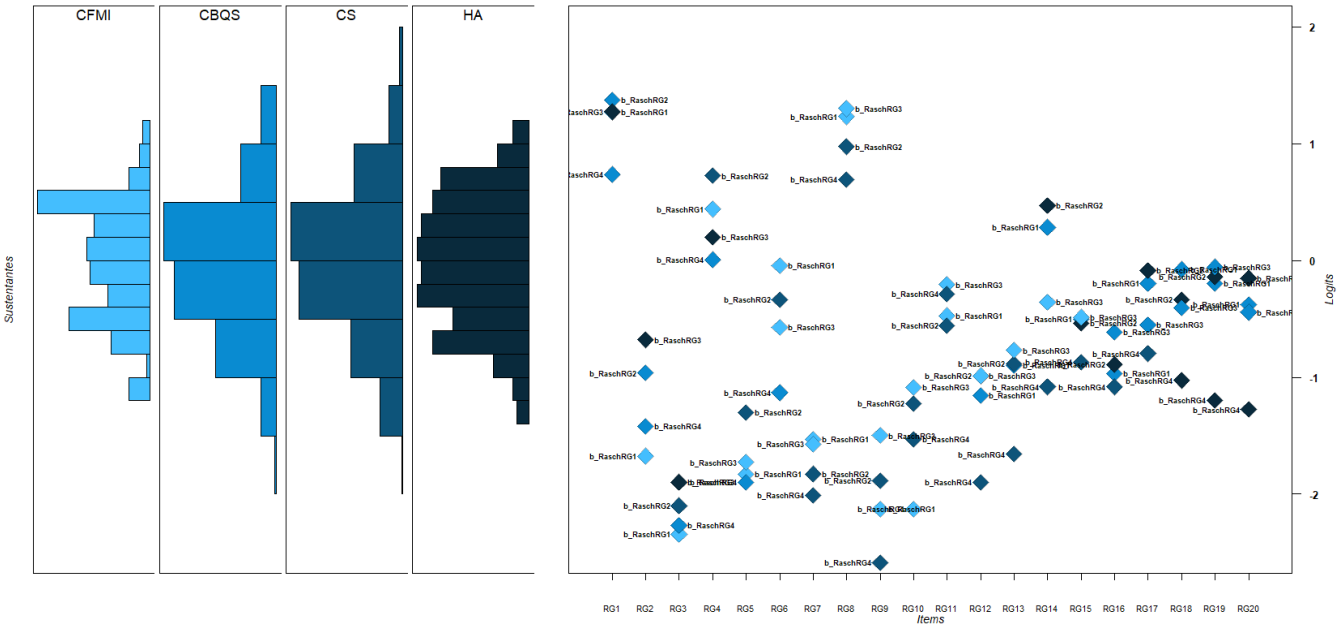
Mapa de Wright del examen de Redacción y Gramática con distribución por área de posgrado



Para poder contrastar también si había diferencias en la dificultad de los reactivos, se calculó el modelo de Rasch permitiendo que los parámetros de dificultad variasen entre cada área de conocimiento. Los resultados se pueden observar en la Figura 11. Se observan diferencias importantes en la dificultad de reactivos como el 1, 5, 7 y 11. Además se aprecia que en general, el parámetro de dificultad suele ser más bajo para el Área HA.

Figura 11.

Mapa de Wright del examen de Redacción y Gramática con dificultad por área de posgrado

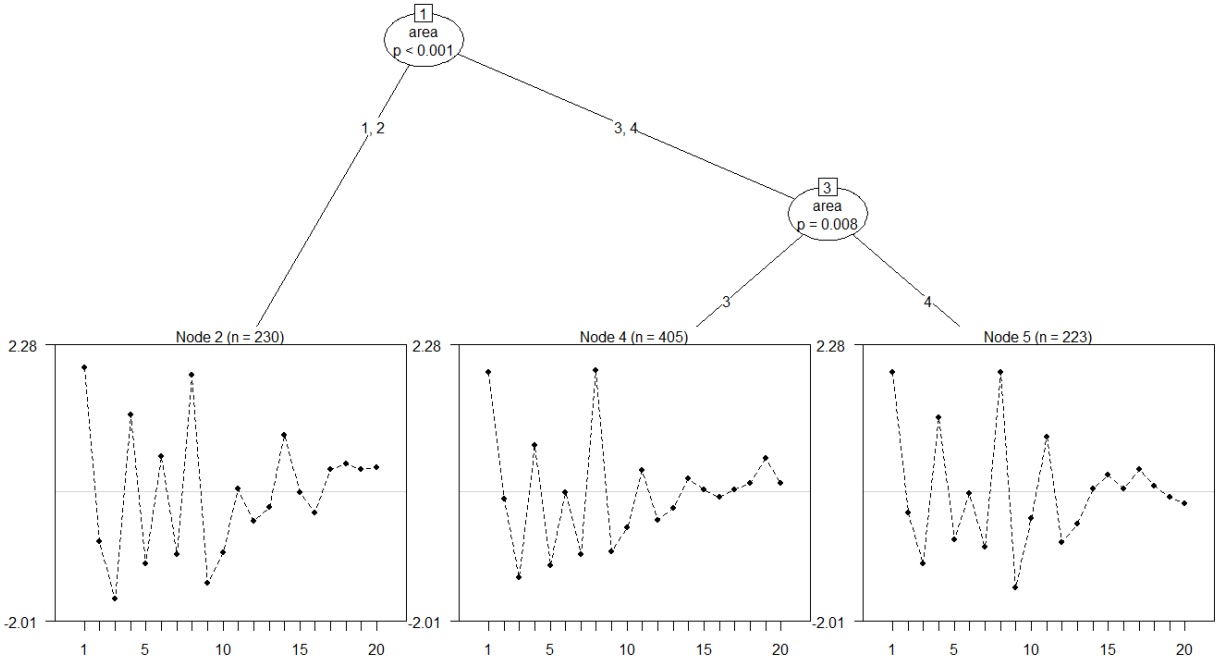


2.3.2 Análisis DIF exploratorio

Para evaluar si la presencia de DIF era específica de las diferencias en las áreas de posgrado, se utilizó el método Raschtree del paquete *psychotree.* Las covariables incluidas fueron la edad, el área del posgrado y el posgrado específico al que aplicaban los sustentantes. Los resultados se pueden observar en la Figura 12.

Figura 12.

Particionamiento de estimación de parámetros de dificultad por área, género y edad con modelo de Rasch y método Raschtree en examen de Redacción y Gramática



Cabe destacar que, pese a que la partición inicial fue similar a lo visto en el examen de comprensión lectora, donde las áreas CFMI y CBQS se agruparon en contraste de las áreas CS y HA, en este caso se observó una segunda partición entre las áreas CS y HA también.

2.3.4 Método de diferencia de logits con modelo de Rasch

En la Tabla 17 se presenta la comparación directa área vs área con el método de diferencia de logits, se observa que los reactivos con mayor presencia de DIF son los reactivos 2, 6, 10 y 14, y los pares de áreas con mayor presencia de DIF fueron las comparativas entre FMI y CS y FMI y HA.

Tabla 19.

Análisis DIF exploratorio del examen de Redacción y Gramática con método de Rasch

Reactivo	FMI y CBQS		FMI y CS		FMI y HA		CBQS y CS		CBQS y HA		CS y HA	
RG1	0.028		0.105		0.063		0.077		0.035		-0.041	
RG2	-0.578	-	-0.887	-	-0.717	-	-0.309		-0.139		0.169	
RG3	-0.113		-0.336		-0.539	-	-0.223		-0.426		-0.203	
RG4	-0.152		0.351		-0.031		0.504	+	0.122		-0.382	
RG5	-0.398		0.008		-0.398		0.406		0.000		-0.406	
RG6	0.425		0.633	+	0.624	+	0.209		0.002		-0.009	
RG7	0.426		0.147		0.013		-0.279		-0.413		-0.134	
RG8	0.387		0.036		0.067		-0.351		-0.320		0.031	
RG9	-0.112		-0.521	-	-0.012		-0.409		0.100		0.509	+
RG10	-0.773	-	-0.931	-	-1.057	--	-0.158		-0.285		-0.126	
RG11	0.220		-0.158		-0.648	-	-0.378		-0.869	-	-0.490	
RG12	-0.035		-0.062		0.275		-0.027		0.310		0.337	
RG13	0.119		-0.023		0.294		-0.142		0.175		0.317	
RG14	-0.049		0.752	+	0.908	+	0.801	+	0.957	+	0.156	
RG15	0.166		0.090		-0.093		-0.076		-0.259		-0.184	
RG16	0.050		-0.249		-0.351		-0.299		-0.401		-0.102	
RG17	0.020		0.465		0.139		0.445		0.119		-0.326	
RG18	0.393		0.437		0.489		0.043		0.095		0.052	
RG19	0.074		-0.031		0.540	+	-0.104		0.466		0.571	+
RG20	-0.098		0.173		0.434		0.271		0.532	+	0.261	

2.3.5 Análisis DIF con método χ^2de Lord

En cuanto al análisis χ^2de Lord presentado en la Tabla 18, los reactivos 2, 11, 12 y 19 son los que presentaron mayor nivel de DIF entre distintas comparaciones. En este caso, las comparaciones con más reactivos con presencia de DIF fueron la comparación entre FMI y HA y CBQS y CS.

Tabla 20.

Análisis DIF exploratorio del examen de Redacción y Gramática con método de χ^2de Lord

Reactivo	FMI y CBQS	FMI y CS	FMI y HA	CBQS y CS	CBQS y HA	CS y HA
RG1	0.206	3.06	7.491 *	3.659	9.779 **	5.292
RG2	10.738 **	17.017 ***	13.893 ***	15.277 ***	10.256 **	2.219
RG3	0.732	2.85	4.799	5.72	10.949 **	2.128
RG4	3.103	3.422	9.714 **	3.254	6.291 *	2.283
RG5	3.422	6.975 *	6.488 *	4.063	7.746 *	4.174
RG6	1.133	1.039	2.402	3.533	3.868	2.015
RG7	1.985	4.188	3.96	5.924	5.927	0.069
RG8	0.394	10.851 **	4.981	20.046 ***	7.557 *	0.476
RG9	2.779	4.464	4.715	7.655 *	2.693	0.472
RG10	0.617	2.376	2.166	4.148	6.606 *	2.542
RG11	30.893 ***	31.561 ***	12.053 **	2.904	4.548	2.341
RG12	2.472	8.562 *	6.568 *	11.507 **	6.302 *	0.764
RG13	2.763	1.889	0.374	11.003 **	2.844	2.331
RG14	3.974	3.289	4.542	4.552	3.503	0.89
RG15	0.691	2.145	1.736	3.474	1.737	0.262
RG16	0.362	4.997	1.195	7.992 *	4.455	0.818
RG17	1.431	3.307	3.748	0.945	4.737	3.425
RG18	0.907	1.974	7.287 *	6.884 *	15.326 ***	4.857
RG19	1.765	42.112 ***	17.754 ***	36.863 ***	14.471 ***	2.338
RG20	5.077	17.794 ***	14.856 ***	8.18 *	5.707	1.696

2.3.6 Comparación de 2 grupos

En cuanto a la comparación entre las áreas CFMI y CBQS como una sola y las áreas CS y HA agrupadas como otra (Tabla 19), se observa los mayores contrastes en la comparación directa de métodos. Por un lado, se observa casi nula presencia de DIF tanto con el método de Mantel-Haenszel (solo reactivos 2, 11 y 14 presentan DIF) como con el de diferencial de Logits (solo el reactivo 14 presenta DIF), mientras que los otros dos métodos señalan la presencia de DIF en una mayor cantidad de reactivos, el método de regresión logística encuentra DIF en 11 reactivos (reactivos 2, 3,4,6,9,10,11,14,16,17,20), y el método de χ^2de Lord señala DIF en 16. Se destaca que el reactivo 14 es el único que muestra DIF en todos los métodos, seguido de los reactivos 2 y 11 que presentan DIF en tres de los cuatro métodos comparados.

Tabla 21.

Análisis DIF exploratorio del examen de Redacción y Gramática con métodos tradicionales

Reactivo	Mantel-Haenszel	Regresión Logística	Diferencia de Logits (Rasch)	χ^2de Lord
RG1	-0.117	0.000	0.072	9.296 **
RG2	1.021 +	0.010 **	-0.372	20.815 ***
RG3	0.893	0.006 *	-0.303	9.260 **
RG4	-0.746	0.005 **	0.328	7.871 *
RG5	-0.337	0.001	0.175	12.270 **
RG6	-0.756	0.005 *	0.332	3.758
RG7	0.590	0.003	-0.222	8.991 *
RG8	0.654	0.004	-0.244	25.374 ***
RG9	0.917	0.006 *	-0.314	6.775 *
RG10	0.815	0.006 *	-0.347	3.772
RG11	1.048 +	0.010 ***	-0.495	1.315
RG12	0.105	0.000	0.050	13.766 ***
RG13	0.067	0.000	-0.016	7.668 *
RG14	-2.020 --	0.031 ***	0.835 +	10.406 **
RG15	0.117	0.000	-0.079	4.157
RG16	0.684	0.004 *	-0.307	7.660 *
RG17	-0.740	0.004 *	0.328	5.465
RG18	-0.296	0.001	0.168	6.248 *
RG19	0.039	0.000	0.094	41.319 ***
RG20	-0.637	0.003 *	0.318	12.797 **

2.4 Métodos paramétricos robustos - IRT

2.4.1 Selección de Modelo

Para el análisis con métodos paramétricos más robustos, se comenzó por determinar qué modelo era el más idóneo para representar los datos. Como se observa en la Tabla 20, todos los modelos comparados presentan un ajuste insuficiente al tener un estadístico M2 con una $p<.05$, sin embargo, sí se puede observar que el modelo de Rasch presenta los estadísticos más bajos con una clara diferencias. Dado que parece que ninguno de los modelos unidimensionales presenta un ajuste satisfactorio, se procedió a utilizar el modelo de 2 parámetros pero con un ajuste multidimensional. Se eligió este modelo por cuestiones de parsimonia y para mantener la consistencia del análisis en ambos exámenes. Además de esto, se consideró el tamaño de muestra que podría tener un efecto importante al estimar los parámetros con modelos más complejos.

Tabla 22.

Comparación de modelos TRI unidimensionales en examen de Redacción y Gramática

Modelos	M2	df	p	RMSEA	CFI
Modelo Rasch	438.586	189	0.000	0.033	0.889
Modelo 2PL	261.004	170	0.000	0.021	0.959
Modelo 3PL	194.955	150	0.008	0.016	0.980

Dado que este examen fue planteado desde un inicio como un examen multidimensional, se procedió a comparar distintos modelos exploratorios desde el unidmensional de dos parámetros hasta modelos exploratorios y confirmatorios de 4 dimensiones. Los resultados se pueden observar en la Tabla 21.

Tabla 23.

Comparación de modelos TRI unidimensionales en examen de Redacción y Gramática

Modelos	M2	df	p	RMSEA	CFI
Modelo unidimensiona	438.586	189	0.000	0.033	0.889
Modelo 2 dimensiones	186.181	151	0.027	0.014	0.984
Modelo 3 dimensiones	142.471	133	0.272	0.008	0.996
Modelo 4 dimensiones	904.563	170	0.000	0.060	0.672
Confirmatorio 4 dimensiones	896.980	170	0.000	0.060	0.675
Confirmatorio parcialmente compensatorio	27,385.619	170	0.000	0.366	0.000

En el contraste directo en la Tabla 21, se presenta como primer modelo estadísticamente significativo bajo el estadístico M2 al modelo de dos dimensiones. Además de esto, el método de razón de verosimilitud mostró que el modelo de dos dimensiones es superior al modelo unidimensional, y se encontró también que la mejoría entre el de 2 dimensiones y el de 3 era marginalmente significativa (Tabla 22). Sin embargo, sí se observa una mejora considerable entre el modelo de 2 dimensiones y el de 4. Otro punto que destacar es que el modelo confirmatorio con las cuatro dimensiones fue el modelo con el peor ajuste, especialmente el modelo parcialmente compensatorio.

Tabla 24.

Comparación de modelos TRI unidimensionales en examen de Redacción y Gramática

Modelos	AIC	BIC	logLik	X2	df	p
Modelo unidimensional	27,984.40	28,188.60	−13,952.20			
Modelo bidimensional	17,177.41	17,340.77	−8,556.71	10,790.99	1046520	1

Dado que este modelo es el que presenta un mejor ajuste a los datos, será el modelo a utilizar en los siguientes análisis basados en la TRI, pero antes de evaluar la presencia de DIF, es importante visualizar la estructura del modelo. Para ello, en la Tabla 23 se presenta la estructura factorial del modelo exploratorio de 2 dimensiones. Como se puede observar, hasta 12 de los 20 reactivos presentan cargas menores a .2 en el segundo factor.

Tabla 25.

Estructura factorial del modelo con dos dimensiones

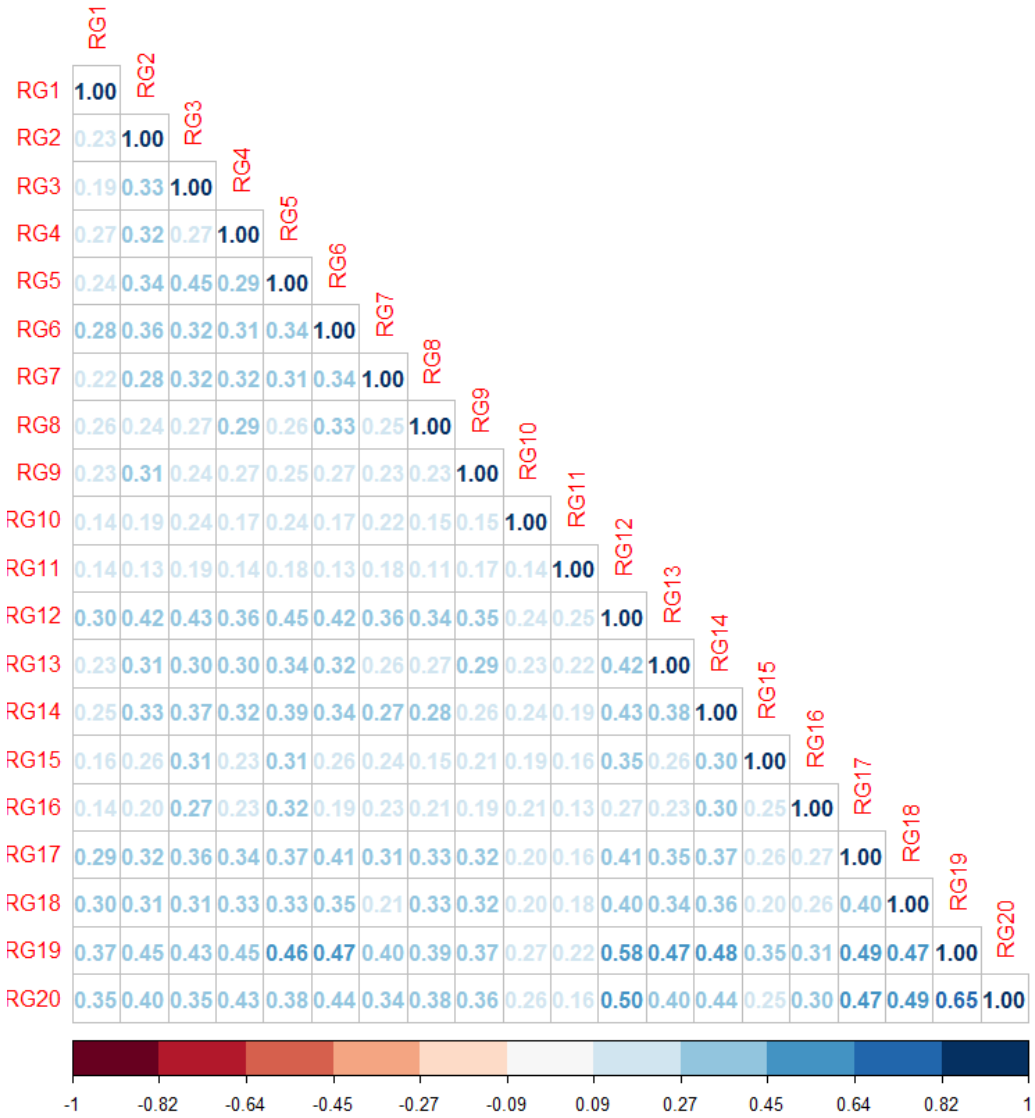
Reactivo	Factor 1	Factor 2
RG1	-0.372	0.040
RG2	-0.344	-0.198
RG3	-0.285	-0.597
RG4	-0.373	-0.048
RG5	-0.286	-0.528
RG6	-0.396	-0.107
RG7	-0.321	-0.268
RG8	-0.373	-0.015
RG9	-0.420	-0.099
RG10	-0.156	-0.216
RG11	-0.105	-0.148
RG12	-0.465	-0.363
RG13	-0.356	-0.194
RG14	-0.325	-0.243
RG15	-0.133	-0.319
RG16	-0.177	-0.233
RG17	-0.417	-0.130
RG18	-0.454	-0.018
RG19	-0.663	-0.194
RG20	-0.663	0.000

2.4.2 Verificación de supuestos

Dado que el nivel de ajuste del modelo multidimensional de 2PL resulta adecuado, es pertinente asumir que el supuesto de monotonicidad es cumplido y que se trata de un modelo no unidimensional. Por ello, únicamente resta verificar en este modelo el supuesto de independencia local. Para ello, se utilizó el estadístico Q3, que permite comparar las correlaciones entre los residuales de los reactivos. Estas correlaciones pueden observarse de forma gráfica en la Figura 13.

Figura 13.

Matriz de valores residuales del test Q3 para evaluar independencia local



De manera sorprendente y contraria a lo observado con el modelo de mejor ajuste en el examen de comprensión lectora, en este caso se observan correlaciones altas entre los residuales de cada uno de los reactivos. Resulta destacable que, al realizar este mismo análisis de residuos con el modelo unidimensional, el patrón observado se asemeja más a lo visto con el modelo del examen de comprensión lectora. Luego de estos resultados, se procedió a contrastar con un segundo método de evaluación del supuesto de independencia local, y de forma contraria, mediante el método X2 de Chen y Thissen (1997a), se obtuvieron valores bajos en todos los cruces de reactivos exceptuando uno (reactivo 18 con reactivo 7). Pese a ello, se ha observado que el método de Q3 tiene mayor sensibilidad para la detección de independencia local, mostrando resultados más confiables en la mayoría de los casos.

A partir de estos resultados contrastantes donde el modelo bidimensional presenta mejor ajuste, pero a su vez no muestra adecuados niveles de independencia local, se procede con cautela a utilizar este modelo con el resto de los análisis y se le compara también con los resultados del modelo unidimensional.

2.4.3 Método de razón de verosimilitud

Para el análisis DIF con el método DIF-Free-Then-DIF, se llevó a cabo un primer análisis con el test de Razón de Verosimilitud utilizando el modelo bidimensional de 2PL. Se observó un primero momento que 6 reactivos no presentaban DIF significativo (reactivos 3, 5, 7, 11, 15 y 16), por lo que se les utilizó como reactivos ancla para un segundo análisis. Los resultados del análisis final que incluye reactivos ancla se pueden observar en la Tabla 24. Se destaca por un lado que bajo este modelo, ninguno de los reactivos presenta DIF significativo. Sin embargo, se observa también que muchos de los reactivos no lograron converger bajo este modelo multidimensional.

Tabla 26.

Análisis DIF con método de Razón de Verosimilitud en examen de Redacción y Gramática con modelo de dos dimensiones

Reactivos	Convergencia	AIC	SABIC	HQ	BIC	X2	df	p
RG1		−3.566	8.005	7.963	27.064	15.566	6	0.016
RG2		−0.037	11.535	11.493	30.593	12.037	6	0.061
RG4		6.232	17.803	17.761	36.862	5.768	6	0.450
RG6		−1.097	10.474	10.432	29.533	13.097	6	0.042
RG8		2.367	13.938	13.896	32.997	9.633	6	0.141
RG9		1.169	12.740	12.699	31.799	10.831	6	0.094
RG10		−15.467	−3.896	−3.938	15.162	27.467	6	0.000
RG12		5.739	17.310	17.268	36.369	6.261	6	0.395
RG13		1.262	12.833	12.791	31.891	10.738	6	0.097
RG14		1.211	12.782	12.740	31.841	10.789	6	0.095
RG17		0.942	12.513	12.471	31.572	11.058	6	0.087
RG18		0.688	12.259	12.218	31.318	11.312	6	0.079
RG19	Convergió	1.280	12.851	12.809	31.909	10.720	6	0.097
RG20		4.851	16.422	16.380	35.481	7.149	6	0.307

En cuanto al modelo unidimensional, el primer modelo sin reactivo ancla mostró los mismos reactivos con ausencia de DIF que el modelo bidimensional (reactivos 3, 5, 7, 11, 15 y 16), Por esto, se tomaron dichos reactivos como reactivos ancla para el siguiente modelo. Los resultados de este segundo análisis se presentan en la Tabla 25. Se observa en primer lugar una mejor convergencia, pues únicamente el reactivo 13 no logró converger bajo este modelo. Sin embargo, se observa también una mayor presencia de DIF, pues en este modelo, 9 reactivos presentaron DIF significativo (reactivos 1, 4, 5, 6, 10, 14, 17 ,18 y 20).

Tabla 27.

Análisis DIF con método de Razón de Verosimilitud en examen de Redacción y Gramática con modelo unidimensional

Reactivos	Convergencia	AIC	SABIC	HQ	BIC	X2	df	p
RG1	Convergió	−8.376	3.195	3.153	22.254	20.376	6	0.002
RG2	Convergió	1.740	13.311	13.269	32.370	10.260	6	0.114
RG4	Convergió	−11.185	0.386	0.345	19.445	23.185	6	0.001
RG5	Convergió	−3.157	8.414	8.372	27.473	15.157	6	0.019
RG6	Convergió	−9.950	1.622	1.580	20.680	21.950	6	0.001
RG8	Convergió	0.120	11.691	11.649	30.750	11.880	6	0.065
RG9	Convergió	3.894	15.465	15.423	34.523	8.106	6	0.230
RG10	Convergió	−2.224	9.347	9.305	28.406	14.224	6	0.027
RG12	Convergió	4.398	15.969	15.927	35.027	7.602	6	0.269
RG13		1.811	13.382	13.341	32.441	10.189	6	0.117
RG14	Convergió	−52.274	−40.703	−40.745	−21.645	64.274	6	0.000
RG17	Convergió	−11.854	−0.282	−0.324	18.776	23.854	6	0.001
RG18	Convergió	−3.732	7.839	7.797	26.897	15.732	6	0.015
RG19	Convergió	0.324	11.895	11.853	30.954	11.676	6	0.070
RG20	Convergió	−8.491	3.080	3.039	22.139	20.491	6	0.002

2.4.4 Análisis DIF con dos grupos

Finalmente, en el análisis correspondiente a comparar las áreas CFMI y CBQS agrupadas contra las áreas CS y HA, se repitió el proceso anterior comparando un modelo bidimensional contra uno unidimensional. Se destaca que, en ambos casos, el primer análisis DIF mostró exactamente los mismos reactivos con ausencia de DIF. Sin embargo, como se puede observar en las tablas X y Y, se observan diferencias importantes entre ambos modelos.

En el modelo bidimensional, se destaca que tres reactivos no lograron converger (reactivos 6, 14 y 17). Se observa en la Tabla 26 también que, entre los reactivos analizados luego de descartar los reactivos ancla, únicamente el reactivo 11 no presenta evidencia de DIF. El reactivo 19 presenta un valor de significancia muy cercano a .05, pero dado el punto de corte establecido, se asume que el reactivo presenta DIF.

Tabla 28.

Análisis DIF con método de Razón de Verosimilitud en examen de Redacción y Gramática con modelo bidimensional - Comparación con 2 grupos

Reactivos	Convergencia	AIC	SABIC	HQ	BIC	X2	df	p
RG1	Convergió	−3.443	0.372	0.388	6.725	7.443	2	0.024
RG4	Convergió	−5.097	−1.281	−1.265	5.072	9.097	2	0.011
RG6		−9.431	−5.615	−5.599	0.738	13.431	2	0.001
RG11	Convergió	0.998	4.814	4.830	11.167	3.002	2	0.223
RG14		−13.296	−9.480	−9.465	−3.128	17.296	2	0.000
RG17		−12.743	−8.928	−8.912	−2.575	16.743	2	0.000
RG18	Convergió	−3.231	0.585	0.601	6.938	7.231	2	0.027
RG19	Convergió	−2.099	1.717	1.733	8.070	6.099	2	0.047
RG20	Convergió	−4.664	−0.848	−0.832	5.505	8.664	2	0.013

Al llevar a cabo el contraste con el modelo unidimensional, se observan dos diferencias principales en la Tabla 27. Por un lado, dado que es un modelo con menor exigencia computacional y menos parámetros, en este caso todos los reactivos logran converger. Sin embargo, la principal diferencia observada es que, a pesar de tener los mismos reactivos ancla, en el resultado final se observa que el reactivo 11 sí es señalado con presencia de DIF bajo este modelo, mientras que el reactivo 19 muestra un valor de *p* mayor al punto de corte, por lo que se asume ausencia de DIF para dicho reactivo.

Tabla 29.

Análisis DIF con método de Razón de Verosimilitud en examen de Redacción y Gramática con modelo unidimensional - Comparación con 2 grupos

Reactivos	Convergencia	AIC	SABIC	HQ	BIC	X2	df	p
RG1	Convergió	−13.998	−10.183	−10.167	−3.830	17.998	2	0.000
RG4	Convergió	−11.576	−7.760	−7.744	−1.407	15.576	2	0.000
RG6	Convergió	−10.815	−6.999	−6.983	−0.646	14.815	2	0.001
RG11	Convergió	−6.625	−2.809	−2.793	3.543	10.625	2	0.005
RG14	Convergió	−55.044	−51.229	−51.213	−44.876	59.044	2	0.000
RG17	Convergió	−10.743	−6.927	−6.911	−0.575	14.743	2	0.001
RG18	Convergió	−4.192	−0.376	−0.360	5.977	8.192	2	0.017
RG19	Convergió	−1.676	2.140	2.155	8.492	5.676	2	0.059
RG20	Convergió	−10.294	−6.479	−6.463	−0.126	14.294	2	0.001

VI. CONCLUSIONES

En el presente estudio se analizó un examen de comprensión lectora y otro de redacción y gramática, diseñado por una IES de la Ciudad de México. El propósito general fue generar evidencias de validez de la interpretación de imparcialidad para los resultados de estos exámenes que serían utilizados como indicadores para el proceso de admisión al posgrado. La interpretación de imparcialidad principal asumida fue que sustentantes a cuatro áreas distintas de los programas de posgrado interpretarían de la misma forma los reactivos del examen; y por ende, tendrían la misma probabilidad de responderlos correctamente, independientemente del área para la que estuvieran aplicando, una vez mediado su nivel de habilidad. Para generar las evidencias de validez, se utilizó una propuesta contemporánea centrada en generar un modelo coherente donde todas las evidencias se alinearan a los mismos supuestos teóricos y métricos.

Para alcanzar dicho propósito, el presente trabajo utilizó la metodología DIF, comenzando por usar procedimientos clásicos para el estudio de DIF, para posteriormente emplear métodos más avanzados alineados a un mismo modelo teórico y empírico para la representación de los datos. Finalmente, se compararon los distintos métodos para detectar DIF, permitiendo una mayor sensibilidad en la detección de reactivos señalados con presencia de DIF a través de las distintas aproximaciones.

La metodología utilizada en este estudio para generar estas evidencias de validez es de suma relevancia, no sólo a nivel teórico-metodológico y psicométrico, como se describirá más adelante; sino también, y principalmente, con respecto a sus implicaciones sociales y éticas. Dentro de los procesos de admisión a las universidades, tanto en estudios de pregrado como de posgrado, el foco de atención se ha llevado cada vez más al uso de herramientas y metodologías que permitan mayor imparcialidad, fomentando a su vez la diversidad de los seleccionados (Reed, 2021).

Ante este creciente interés por la justicia y la imparcialidad (AERA et al., 2014), los exámenes que se sigan utilizando como indicadores dentro de estos procesos no pueden quedarse atrás, por lo que requieren presentar evidencias sólidas en proporción al impacto de las interpretaciones y usos que se hagan a partir de los puntajes de estos exámenes (Messick, 1989). Para lograrlo, no solamente se requiere utilizar técnicas sofisticadas estadísticamente hablando (Baker, 2001), sino que se requiere que las múltiples herramientas utilizadas para generar evidencia de las distintas interpretaciones sigan un hilo conductor que mantenga suficiente coherencia teórica y metodológica, no solamente en términos psicométricos (Kane, 2006), sino también en relación con las características específicas de cada examen y de lo que pretendan medir, especialmente al tratarse de constructos tan complejos como las habilidades verbales (Leighton y Gierl, 2007).

Tomando todos estos puntos en cuenta, este estudio permitió ejemplificar la importancia de diseñar e implementar una postura metodológica sólida y coherente en el análisis DIF. Al utilizar la Teoría de Respuesta al Ítem como hilo conductor desde el inicio, se pudieron generar los modelos más adecuados para representar los datos en ambos exámenes, para después utilizar los que mejor representaban los datos para la detección de DIF en los reactivos, complementando estos análisis con otros métodos no paramétricos. Además, se pudo contrastar qué manera de particionar los grupos explicaba mejor la presencia de DIF.

De manera general, se observó que la presencia de DIF resulta más significativa en la comparación entre las áreas CFMI y CBQS contra las áreas CS y HA, observando una mayor diferencia significativa entre el área HA y el resto de las áreas. Debido a esto y a lo hipotetizado previamente por la DEE -que las diferencias principales estarían entre estos dos grupos-, se considera que los análisis con mayor representatividad y relevancia son los análisis DIF con dos grupos.

De manera consistente, tanto en los análisis área contra área como en los análisis de FM y CBQS como un grupo y CS y HA como otro, se encontró un nivel alto de consistencia en cuanto a la presencia de DIF para ciertos reactivos. En el examen de comprensión lectora, los reactivos con mayor presencia de DIF en las distintas aproximaciones utilizadas fueron los reactivos 4, 5 y 6, y en menor medida los reactivos 7, 9 y 10. Por su parte, en el examen de redacción y gramática, los reactivos con mayor presencia de DIF fueron los reactivos 2, 6 y 14, y en menor medida los reactivos 10 y 11.

En cuanto al mejor modelo a utilizar para cada examen, se encontró que el examen de comprensión lectora es mejor representado por un modelo unidimensional de 2 parámetros, presentando suficiente evidencia del cumplimiento de supuestos de este modelo, teniendo además un nivel de ajuste adecuado. Este hallazgo es concordante con el diseño original del examen que se planteó desde un inicio como unidimensional.

Por su parte, el examen de redacción y gramática es mejor representado por un modelo de 2 parámetros multidimensional compensatorio. Si bien en su tabla de especificaciones original, el examen se planteó con una estructura de cuatro dimensiones distintas con sus respectivos reactivos, en la práctica, dicho modelo explicativo no ajustaba de manera adecuada a los datos, así como tampoco lo hizo el modelo unidimensional. En su lugar, el modelo utilizado finalmente fue un modelo de 2 dimensiones compensatorias. Esto indicaría que, para la resolución de cada reactivo, el sustentante requiere echar mano en mayor o menor medida para cada reactivo de dos habilidades diferenciables. En otras palabras, el cúmulo de reactivos utilizados dentro de este examen requieren de más de una habilidad para su resolución.

Estos hallazgos son de suma relevancia, especialmente porque los reactivos señalados con presencia de DIF varían entre estas dos aproximaciones. Dado que la presencia de DIF también puede ser definida como varianza sistemática irrelevante al constructo (en este caso, a la

habilidad), el hecho de que para estos reactivos se requiere de más de una habilidad, podría indicar que parte de la presencia de DIF es en realidad correspondiente a esta segunda dimensión no considerada en otro tipo de modelos (De Boeck et al., 2011).

Pese a que el modelo con mejor ajuste de los analizados dentro del marco de la TRI fue el modelo de 2 parámetros con dos dimensiones compensatorias, este no careció de dificultades. En particular, aunque su ajuste a los datos fue adecuado, el supuesto de independencia local no se cumplió, ya que se observó una correlación alta entre los residuales de distintos pares de reactivos. Además, al observar las cargas factoriales, se pudo constatar que, para el segundo factor, muchos de los reactivos presentaban cargas muy bajas.

Estos dos resultados técnicos se traducen en una implicación práctica sumamente compleja: la interpretabilidad de los resultados. Si bien se pudo probar que el examen de redacción y gramática no cumple con el supuesto de unidimensionalidad, también se presentó evidencia sobre el pobre ajuste del modelo de cuatro dimensiones propuesto por la DEE de la IES. Por ello, al utilizar el modelo exploratorio de dos dimensiones, interpretar qué representa cada una de estas dimensiones se puede convertir en un desafío importante, en especial dado que para el presente estudio no se tuvo acceso a los reactivos específicos del examen, solamente a la tabla de especificaciones. Por todo ello, más adelante al mencionar las limitaciones y recomendaciones se discutirán algunas posibles alternativas para solventar esta problemática.

En términos generales, y dados los resultados principales obtenidos que recién se discutieron, el presente estudio pudo cumplir de manera parcial los objetivos planteados en un inicio. En cuanto al objetivo principal, sí se generaron evidencias de validez sobre la interpretación de imparcialidad para la gran mayoría de los reactivos de ambos exámenes. Además, el uso de distintas técnicas para detectar DIF permitió generar otras evidencias más robustas en ese sentido. Para el examen de comprensión lectora, 3 reactivos principalmente no presentan evidencia

suficiente para sustentar la interpretación de imparcialidad, mientras que, para el examen de redacción y gramática, fueron 3 reactivos con una mayor presencia de DIF. Por lo que estos reactivos requieren una revisión a profundidad para determinar las fuentes de sesgos y tomar una decisión sobre su eliminación, modificación o interpretación.

Si bien los objetivos anteriormente descritos se alcanzaron de manera satisfactoria, un objetivo específico de esta investigación que se cumplió solamente de forma parcial fue el referente a determinar el modelo estadístico y la metodología DIF más adecuada para las condiciones específicas de cada examen. En particular, para el examen de redacción y gramática, dada la complejidad del modelo que presentó el mejor nivel de ajuste y las limitaciones que tuvo y que ya se describieron más arriba, las evidencias generadas requieren interpretarse con cautela.

No obstante la limitación señalada, que más adelante se discutirá, el nivel de cumplimiento de los objetivos es suficiente como para permitir extraer algunas conclusiones importantes. En primer lugar, las evidencias presentadas permiten identificar con un alto nivel de confianza los reactivos con mayor presencia de DIF a través de los distintos métodos utilizados; lo que permitirá, como un segundo paso -más allá de los alcances del presente estudio-, llevar a cabo una revisión más cualitativa sobre el contenido de los reactivos para determinar la mejor forma de proceder con cada uno de ellos.

En segundo lugar, en términos de dimensionalidad, los resultados son consistentes con lo observado en otros estudios, pues el examen de comprensión lectora presentó una única dimensión, como lo describió Thirakunkovit (2018). Por otra parte, el examen de redacción y gramática presentó algunas dificultades esperadas; como ya se describió en la sección del marco teórico, valorar la habilidad de escritura y de redacción a partir de reactivos de opción múltiple puede conllevar serias limitaciones y desventajas. Además, dado el insuficiente nivel de ajuste de un

modelo unidimensional para representar este examen, queda clara la importancia de repensar la manera en que se evalúa este constructo con este tipo de examen.

En términos metodológicos, a partir de los resultados obtenidos se concluye que el uso de distintas técnicas y metodologías para detectar DIF es esencial para reducir errores de tipo I; como se pudo observar, si bien hubo consistencia respecto a los principales reactivos señalados con presencia de DIF a través de las distintas metodologías algunos reactivos fueron señalados solo por ciertos métodos. En caso de haber utilizado solo uno de estos métodos, se pudo haber señalado reactivos que no presentan DIF al contrastarlos con otras metodologías; o, por el contrario, se podrían haber tomado reactivos como imparciales aún con presencia de DIF no detectada.

Sin embargo, el uso de distintas metodologías debe seguir un razonamiento lógico e intencionado; debido a que es necesario comprender y tomar en cuenta los distintos supuestos subyacentes a cada metodología utilizada, para con ello, generar interpretaciones de los resultados pertinentes y alineadas no solamente con un marco teórico coherente, sino también con las necesidades y características de cada examen, así como de los usos e interpretaciones previstos (Kane, 2021).

Dentro de los resultados obtenidos se destaca el uso de metodologías recursivas de carácter exploratorio, que permiten respaldar o descartar de manera empírica los supuestos establecidos a nivel teórico o técnico con respecto a qué grupos se espera presenten diferencias significativas. En este caso, tanto el método no paramétrico (regresión logística) como el método TRI (Raschtree) aportaron evidencia de validez respecto a los grupos para analizar la presencia de DIF. Es importante señalar que, en el caso del método de Raschtree, al basarse en un modelo Rasch, que como se vio en los análisis posteriores, no presenta un ajuste adecuado a los datos, los resultados pudieran incluir en sí mismos un nivel de sesgo importante. De ahí que se reitera la importancia de contrastarlo con metodologías más acordes a la estructura y características específicas de los datos.

VII. DISCUSIÓN

Más allá de las implicaciones teórico-metodológicas señaladas, el presente estudio permite también abrir el espacio de discusión sobre implicaciones educativas y de carácter práctico a partir de los resultados obtenidos. Como se mencionó al inicio de las conclusiones, la relevancia del uso de estas metodologías en exámenes de alto impacto queda de manifiesto en el presente estudio. Se pudo detectar a partir de distintas metodologías que varios de los reactivos de los dos exámenes analizados presentan DIF considerable. Obviar este hecho podría llegar a tener implicaciones serias en términos de justicia e imparcialidad (AERA et al., 2014); personas aplicantes a áreas como la CFMI y II de los posgrados de la IES habrían sido perjudicadas de forma directa al equiparar sus resultados en estos exámenes con los de personas aplicantes a las otras áreas, sin tomar en cuenta los sesgos que algunos de los reactivos presentaron. Si bien estos exámenes no serán los únicos medios utilizados para determinar quiénes acceden a los programas educativos, el simple hecho de ser parte del proceso crea la necesidad, de un posicionamiento ético y de justicia, de generar evidencias sólidas teórica y empíricamente que respalden los usos e interpretaciones que se harán de los resultados de ese tipo de exámenes.

Aunque este estudio permite generar un panorama amplio con respecto al sesgo de medición a nivel de reactivo de los exámenes de comprensión lectora y de redacción y gramática, es importante interpretar los resultados con cautela y considerar que el presente trabajo contó con limitaciones importantes. En primer lugar, y de mayor relevancia, una limitación fue el tamaño de muestra. Si bien, al contar con más de mil sustentantes, es posible realizar la mayoría de los análisis no paramétricos e incluso algunos de los modelos IRT (Bolt, 2002), al tomar en consideración que la base de datos se termina dividiendo hasta en cuatro partes (una para cada área de posgrado), cada una con un tamaño de muestra distinto, los resultados se vuelven menos robustos.

Al contar con solo 147 aplicantes del área CFMI y 227 del área HA, los resultados de los análisis DIF realizados, área contra área, han de ser tomados con cautela, dado que el tamaño de muestra no es el más indicado. Esta situación podría, de manera parcial, ser la explicación del por qué el análisis DIF con el método de Razón de Verosimilitud con cuatro grupos no logró converger en muchos de los reactivos. En relación con este último punto, las principales limitaciones se observaron con respecto al examen de redacción y gramática. Dada la complejidad surgida desde el diseño de la tabla de especificaciones con cuatro dimensiones claramente diferenciadas pero que a su vez implicaban procesos comunes, el tamaño de muestra se volvió una limitante al momento de ejecutar los análisis más robustos con modelos multidimensionales (Kose y Demirtasli, 2012).

El presente estudio permitió determinar que ni el modelo confirmatorio basado en la estructura de la tabla de especificaciones, ni un modelo unidimensional son suficientes para representar los datos de este tipo de exámenes. Sin embargo, no fue posible determinar un modelo con suficiente nivel de ajuste y que a su vez fuera fácilmente interpretable. Además, muchos de los análisis DIF que utilizaron los modelos multidimensionales no convergieron como se esperaba, por lo que las evidencias generadas sobre la presencia de DIF no terminaron por ser representativas ni acordes con el modelo que mejor explicaba los datos.

Una última limitación no menos importante es el aislamiento del presente estudio con respecto al resto de evidencias de usos e interpretaciones necesarios para estos exámenes. Si bien hay estudios previos que aportan evidencias importantes (Dirección de Evaluación Educativa, 2017), estos siguen un marco teórico psicométrico distinto, por lo que la alineación aquí presentada entre los distintos análisis queda limitada a una única interpretación de los resultados de los exámenes.

No obstante las limitaciones, el presente estudio sirve como un punto de partida que deja de manifiesto la importancia de alinear las evidencias de los usos e interpretaciones a un marco

teórico-psicométrico que permita representar los datos con un nivel de ajuste adecuado. Dicho esto, el trabajo por desarrollar hacia adelante es amplio.

En primer lugar, se sugiere replicar los análisis aquí realizados con una muestra más amplia para las cuatro áreas de conocimientos del posgrado, lo que permitirá generar evidencias más sólidas y robustas. Esto permitirá también verificar que los reactivos señalados con DIF se mantengan de forma consistente a través de ambas muestras, al menos en el examen de comprensión lectora. En cuanto al examen de redacción y gramática, el presente estudio no alcanzó a generar suficientes evidencias de validez de la interpretación de imparcialidad. Para lograrlo, se recomienda utilizar otras técnicas y metodologías que permitan generar las evidencias psicométricas de dicha interpretación.

Una alternativa pertinente podría ser el uso de modelos TRI más complejos que se ajusten a la complejidad misma del examen. Por un lado, el modelo de reactivos localmente dependientes (LID por sus siglas en inglés) de Ip (2010), podría ser una alternativa a los modelos multidimensionales, permitiendo la interrelación entre los reactivos sin la necesidad de recurrir a dimensiones adicionales para explicar los patrones de respuesta. La ventaja principal de este modelo sería reducir los requisitos computacionales necesarios para modelar los datos, permitiendo así generar evidencias, aún con tamaños de muestra menores a los sugeridos para otros análisis. Además, se haría más sencilla la interpretación de los resultados.

Por otro lado, una opción interesante es la propuesta de Molenaar (2015), de un modelo heteroscedástico que permita tomar en consideración la propia complejidad de cada reactivo, que pudiera estar detrás de la presencia de DIF de algunos de los reactivos. El modelo heteroscedástico propuesto es especialmente útil en la evaluación de exámenes como los analizados, ya que toma en consideración la complejidad inherente a cada reactivo. Este enfoque se centra en la idea de que la

varianza de los errores de medición puede variar entre los diferentes reactivos, lo que podría ser la causa de la presencia de DIF en algunos de ellos.

El modelo heteroscedástico de Molenaar (2015) permite analizar cómo los errores de medición se distribuyen de manera desigual entre los diferentes reactivos, lo que podría explicar por qué algunos reactivos funcionan de manera diferente para distintos grupos de individuos. Una de las ventajas del enfoque es que ofrece una perspectiva más realista y flexible de cómo los errores de medición podrían estar influyendo en el desempeño de los examinados. Esto permite mejorar la precisión de las estimaciones de habilidad, al tomar en cuenta la complejidad de cada reactivo y ajustar las estimaciones en función de las diferencias en la varianza de los errores de medición.

Una tercera opción que permitiría contrastar la dimensionalidad del examen a partir de un marco teórico más robusto en términos de dimensionalidad sería utilizar un modelo MIMIC (Modelo de múltiples indicadores y múltiples causas) a partir de la TCT y de las metodologías basados en análisis factorial y ecuaciones estructurales (Finch, 2005). Diversos estudios han demostrado que ambas aproximaciones (IRT y MIMIC) pueden ser análogas matemáticamente (Glockner-Rist y Hoijtink, 2003; Lance et al., 2010). Es por esto que, modelar los datos a partir de un modelo MIMIC permitirá aproximarse al tema de la dimensionalidad desde otro marco teórico más utilizado para propuestas multidimensionales, y aún así permitirá analizar la presencia de DIF.

En suma, evidencias adicionales permitirán determinar el mejor modelo a utilizar, y con ello, confirmar qué reactivos requieren mayor atención con respecto a su funcionamiento diferencial. Sin embargo, el trabajo no ha de quedar ahí. Con los resultados aquí obtenidos queda de manifiesto, tal como lo especifican los estándares (AERA et al., 2014), el nivel de rigurosidad de las evidencias generadas debe ser proporcional al nivel de impacto que éstas pueden tener en las vidas de las personas. Por ello, dada la trascendencia que implica ingresar o no a un programa de

posgrado en una universidad con alto prestigio, las herramientas utilizadas en el proceso de selección deben contar con suficientes evidencias que estén a la altura de la evaluación.

Una vez identificados los reactivos con presencia de DIF, más allá de las decisiones que se tomen con respecto a cada uno, ya sea modificar, eliminar o mediar estadísticamente, aún quedarían pasos importantes a tomar en cuenta. Por un lado, dado que estos exámenes serán un criterio de selección importante, establecer puntos de corte adecuados será un paso fundamental para poder utilizarlos con este propósito (Cizek y Bunch, 2006). Y en ese sentido, no está de más señalar nuevamente la importancia de realizar este procedimiento a partir de un marco teórico coherente con el resto de los análisis aquí presentados.

Por otro lado, dado que la validez no es un tema finito, sino que requiere de un trabajo constante (AERA et al., 2014), será importante también volver a valorar la relevancia práctica de estos exámenes que permita respaldar su relevancia como herramientas de selección. Por un lado, será importante valorar su utilidad predictiva con respecto al nivel de logro de las personas admitidas en el posgrado. Además, sería pertinente contrastar sus resultados con otros medios para valorar la comprensión lectora, redacción y gramática de los sustentantes, especialmente dada la complejidad de estos constructos (Grabe y Stoller, 2019; Hyland, 2021). Inclusive, dada esta misma complejidad, valdría la pena pensar en alternativas de medición para estos constructos. Por un lado, moverse del terreno de los reactivos de opción múltiple hacia respuesta construida permitiría generar evidencias cercanas a contextos reales, y con mayor relevancia práctica (Weir, 2005).

En la actualidad existen modelos teóricos para valorar exámenes más complejos que los de opción múltiple (Gebril, 2018; Lievens y Sackett, 2006), por lo que considerar dichas alternativas podría permitir generar evidencias más robustas, no del funcionamiento de los exámenes, sino de las habilidades reales de los sustentantes. Sería por lo menos pertinente generar este contraste

para determinar qué pruebas o estrategias resultan más relevantes y útiles, tanto en términos de costo-beneficio como de justicia e imparcialidad.

En última instancia, todo examen de alto impacto requiere de evidencias proporcionales a la complejidad de sus usos e interpretaciones (Kane, 2021), y en el caso de los exámenes de habilidades verbales, la complejidad es extensa. Dentro de los estándares para la evaluación de la lectura y escritura (IRA y NCTE, 2009) se señala claramente que es imposible producir exámenes de habilidades verbales que estén libres de sesgos, pero pese a esa inevitabilidad, cuando se requiere usar exámenes, es indispensable controlar tantos sesgos como sea posible.

VIII. REFERENCIAS

AERA, APA, & NCME. (2014). Estándares para Pruebas Educativas y Psicológicas. En Trans. (M. Lieve (Ed.), *Estándares para Pruebas Educativas y Psicológicas* ((Original). American Educational Research Association. (Original. https://doi.org/10.2307/j.ctvr43hg2

Alderson, C. J., Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Almarabheh, A., Shehata, M. H., Ismaeel, A., Atwa, H., & Jaradat, A. (2022). Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study. *Frontiers in Medicine*, *9*. https://doi.org/10.3389/fmed.2022.971926

Álvarez Montero, F., MojardínHeráldez, A., & AudeloLópez, C. (2014). Criterios e Instrumentos para la Admisiónen los Estudios de Doctorado. *Electronic Journal of Research in Educational Psychology*, *12*(3), 853–886.

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Citeseer.

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A Comparative Analysis of the Efficacy of Differential Item Functioning Detection for Dichotomously Scored Items among Logistic Regression, SIBTEST and Raschtree Methods. *Procedia - Social and Behavioral Sciences*, *191*, 21–25. https://doi.org/10.1016/j.sbspro.2015.04.664

Assessment, I. J. T. F. on, Association, I. R., & English, N. C. of T. of. (2009). *Standards for the assessment of reading and writing*. International Reading Assoc.

Backhoff, E., & Contreras Roldán, S. (2014). “Corrupción de la medida” e inflación de los resultados de ENLACE. *Revista mexicana de investigación educativa*, *19*(63), 1267–1283.

Baehman, L. F. (2007). *What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment* (pp. 41–71).

Baines, L. A., & Stanley, G. K. (2005). High-Stakes Hustle: Public Schools and the New Billion Dollar Accountability. *The Educational Forum*, *69*(1), 8–15. https://doi.org/10.1080/00131720408984660

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bastedo, M. (2021). Holistic Admissions as a Global Phenomenon. En H. Eggins, A. Smolentseva, & H. de Wit (Eds.), *Higher Education in the Next Decade* (pp. 91–114). BRILL. https://doi.org/10.1163/9789004462717_006

Blazer, C. (2011). Unintended Consequences of High-Stakes Testing. Information Capsule. Volume 1008. *Research Services, Miami-Dade County Public Schools*.

Bock, R. D., & Moustaki, I. (2006). *Item Response Theory in a General Framework* (C. R. Rao & S. Sinharay, Eds.; Vol. 26, pp. 469–513). https://doi.org/10.1016/S0169-7161(06)26015-2

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education*, *15*(2), 113–141. https://doi.org/10.1207/S15324818AME1502_01

Bond, L., Moss, P., & Carr, P. (1996). Fairness in large-scale performance assessment. En *Technical issues in large-scale performance assessment* (pp. 117–140). National Center for Education Statistics Washington, DC.

Bonifay, W., & Cai, L. (2017). On the Complexity of Item Response Theory Models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Briggs, D. C., & Wilson, M. (2003). An Introduction to Multidimensional Measurement using Rasch Models. *Journal of Applied Measurement*, *4*(1), 87–100.

Burgoyne, A. P., Mashburn, C. A., & Engle, R. W. (2021). Reducing adverse impact in high-stakes testing. *Intelligence*, *87*, 101561. https://doi.org/10.1016/j.intell.2021.101561

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, *11*(3). https://doi.org/10.1002/wics.1460

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. En F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing* (pp. 32–70). Cambridge University Press.

Chapelle, C. A. (2013). Conceptions of validity. En G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.

Chen, W.-H., & Thissen, D. (1997a). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Cheng, C.-P., Chen, C.-C., & Shih, C.-L. (2020). An exploratory strategy to identify and define sources of differential item functioning. *Applied Psychological Measurement*, *44*(7–8), 548–560. https://doi.org/https://doi.org/10.1177/0146621620931190

Cho, S.-J., Suh, Y., & Lee, W. (2016). After Differential Item Functioning Is Detected. *Applied Psychological Measurement*, *40*(8), 573–591. https://doi.org/10.1177/0146621616664304

Cizek, G. J., & Bunch, M. B. (2006). Standard setting. *Handbook of test development*, 225–258.

Cronbach, L. J. (2013). Five perspectives on validity argument. En *Test validity* (pp. 3–17). Routledge.

Davies, A. (1990). *Principles of Language Testing.* Blackwell.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory Secondary Dimension Modeling of Latent Differential Item Functioning. *Applied Psychological Measurement*, *35*(8).

Finch, H. (2005). The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, *29*(4), 278–295.

García Medina, A. M., Martínez Rizo, F., & Cordero Arrollo, 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.

Gebril, A. (2018). Integrated-Skills Assessment. En *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0544

Glockner-Rist, A., & Hoijtink, H. (2003). The Best of Both Worlds: Factor Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(4), 544–565. https://doi.org/10.1207/S15328007SEM1004_4

Grabe, W., & Stoller, F. L. (2019). *Teaching and researching reading*. Routledge.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical care*, *44*(11), S182–S188.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hatch, E., & Lazaraton A. (1997). *The Research Manual: Design and Statistics for Applied Linguistics.* Heinle and Heinle Publishers.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, *103*(2), 219–230. https://doi.org/10.1007/s11205-011-9843-4

Hyland, K. (2021). *Teaching and researching writing*. Routledge.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 395–416. https://doi.org/10.1348/000711009X466835

Jiménez Catalán, R. M. (2002). El concepto de competencia léxica en los estudios de aprendizaje y enseñanza de segundas lenguas. *Atlantis*, *24*(1), 149–162.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Juarros, M. F. (2006). ¿EDUCACIÓN SUPERIOR COMO DERECHO O COMO PRIVILEGIO? *Andamios, 3*(5), 69–90.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger Publishers.

Kane, M. T. (2021). Articulating a validity argument. En *The Routledge handbook of language testing* (pp. 32–47). Routledge.

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of Unidimensional and Multidimensional Models Based on Item Response Theory in Terms of Both Variables of Test Length and Sample Size. *Procedia - Social and Behavioral Sciences*, *46*, 135–140. https://doi.org/10.1016/j.sbspro.2012.05.082

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. *Educational and Psychological Measurement*, *65*(6), 935–953. https://doi.org/10.1177/0013164405275668

Lance, C. E., Lance, C. E., & Vandenberg, R. J. (2010). The partial revival of a dead horse? Comparing classical test theory and item response theory. En *Statistical and methodological myths and urban legends* (pp. 57–80). Routledge.

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*, *3*(9), 40–55. https://doi.org/10.1016/S2007-5057(14)72724-3

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive Diagnostic Assessment for Education* (J. Leighton & M. Gierl, Eds.). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Lin, W.-L., & Yao, G. (2014). Predictive Validity. En *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5020–5021). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_2241

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdova, N., Jiménez, D., Blanco, M. P., Ortega, G., & Lemarié, J. (2021). The Position of Distractors in Multiple-Choice Test Items: The Strongest Precede the Weakest. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.731763

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274. https://doi.org/https://doi.org/10.1177/0013164412453841

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.

Makransky, G., Havmose, P., Vang, M. L., Andersen, T. E., & Nielsen, T. (2017). The predictive validity of using admissions testing and multiple mini-interviews in undergraduate university admissions. *Higher Education Research & Development*, *36*(5), 1003–1016. https://doi.org/10.1080/07294360.2016.1263832

Marta Ferreyra, M., Avitabile, C., Botero Álvarez, J., Haimovich Paz Sergio Urzúa, F., & Humano, D. (2017). *Momento decisivo La educación superior en América Latina y el Caribe Resumen*.

Martínez González, A., Contreras Michel, N. S., Londoño Cárdenas, M. J., & Manzano Patiño, A. P. (2017). Informe del valor predictivo del Examen de Ingreso al Programa Único de Especializaciones en Economía, Generaciones 2015 y 2016.

Maydeu-Olivares, A. (2014). Evaluating the fit of IRT models. En *Handbook of item response theory modeling* (pp. 129–145). Routledge.

Maydeu-Olivares, A., & Joe, H. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement*, *22*(4), 357–367.

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, Justice and Language Assessment*. Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Wiley-Blackwell.

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

Mendoza Vega, J. B., & Corona Burch, M. J. (2018). Funcionamiento diferencial de los reactivos de Excale 03-2010 de Español en estudiantes hablantes de una lengua indígena. Análisis DIF, minería de texto y análisis cualitativo de especiFicaciones. En A. Navarrete Zumárraga & L. López Pérez (Eds.), *Tendencias de Investigación e Innovación en Evaluación Educativa. Memoria del Simposio* (pp. 100–112). Fondo Sectorial de Investigación para la Evaluación de la Educación CONACyT-INEE.

Menken, K. (2009). NO CHILD LEFT BEHIND AND ITS EFFECTS ON LANGUAGE POLICY. *Annual Review of Applied Linguistics*, *29*, 103–117. https://doi.org/10.1017/S0267190509090096

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan Publishing Company.

Messick, S. (1993). FOUNDATIONS OF VALIDITY: MEANING AND CONSEQUENCES IN PSYCHOLOGICAL ASSESSMENT. *ETS Research Report Series*, *1993*(2), i–18. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy / Pedagogický casopis*, *3*(1), 82–100. https://doi.org/10.2478/v10159-012-0004-x

Mislevy, R. J., & Yin, C. (2013). Evidence-centered design in language testing. En *The Routledge handbook of language testing* (pp. 222–236). Routledge.

Molenaar, D. (2015). Heteroscedastic Latent Trait Models for Dichotomous Data. *Psychometrika*, *80*(3), 625–644. https://doi.org/10.1007/s11336-014-9406-0

Morales Ardaya, F. (2004). Evaluar la escritura, sí... Pero ¿Qué y cómo evaluar?. *Acción Pedagógica*, *13*(1), 38–48.

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, *4*(2), 199–203. https://doi.org/10.1002/wics.199

OECD. (2019). *PISA 2018 Results: What Students Know and Can Do: Vol. Volume I*. OECD Publishing. . https://doi.org/10.1787/5f07c754-en

Oleas-Falconí, D. Z., & Mosquera-Endara, M. del R. (2022). La meritocracia y su efecto en la educación superior. *IUSTITIA SOCIALIS*, *7*(2), 1358. https://doi.org/10.35381/racji.v7i2.2387

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

Osterlind, S. J. (1983). *Test item bias* (Número 30). Sage.

Penfield, R. D. (2016). Fairness in test scoring. En N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 71–92). Routledge.

Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. *Handbook of Statistics*, *26*, 125–167. https://doi.org/10.1016/S0169-7161(06)26005-X

Purpura, J. E. (2012). Assessment of Grammar. En *The Encyclopedia of Applied Linguistics*. Wiley. https://doi.org/10.1002/9781405198431.wbeal0045

Reed, H. (2021). Diversity requires an admissions process overhaul. *Journal of the American Academy of Physician Assistants*, *34*(6), 11–12. https://doi.org/10.1097/01.JAA.0000750984.86427.a3

Reeves, R. V., & Halikias, D. (2017). *Race gaps in SAT scores highlight inequality and hinder upward mobility*.

Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. En *Handbook of item response theory modeling* (pp. 31–58). Routledge.

Rivas, A., & Scasso, M. G. (2021). Low stakes, high risks: the problem of intertemporal validity of PISA in Latin America. *Journal of Education Policy*, *36*(2), 279–302. https://doi.org/10.1080/02680939.2019.1696987

Rogers, H. J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Sánchez-Mendiola, M., & Delgado-Maldonado, L. (2017). Exámenes de alto impacto: implicaciones educativas. *Investigación en Educación Médica*, *6*(21), 52–62. https://doi.org/10.1016/j.riem.2016.12.001

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of Approximation Techniques for Detecting Item Bias. *Journal of Educational Measurement*, *22*(2), 77–105. https://doi.org/10.1111/j.1745-3984.1985.tb01050.x

Silva, M. (2020). La dimensión pedagógica de la equidad en educación superior. *Education Policy Analysis Archives*, *28*, 46. https://doi.org/10.14507/epaa.28.5039

Sternberg, R. J. (2000). Standardized Minds: The High Price of America's Testing Culture and What We Can Do To Change It. *NASSP Bulletin*, *84*(616), 118–121. https://doi.org/10.1177/019263650008461617

Sternberg, R. J. (2010). *College Admissions for the 21st Century*. Harvard University Press.

Stobart, G. (2003). The Impact of Assessment: Intended and unintended consequences. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 139–140. https://doi.org/10.1080/0969594032000121243

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. https://doi.org/10.1080/0969594X.2012.639191

The College Board. (2017). *SAT. Suite of Assessments Technical Manual: Characteristics of the SAT*. The College Board.

Thirakunkovit, S. (2018). The Notions of Language Proficiency and Language Dimensionality. *PASAA: Journal of Language Teaching and Learning in Thailand*, *55*, 219–236.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On Variable Importance in Linear Regression. *Social Indicators Research*, *45*(1/3), 253–275. https://doi.org/10.1023/A:1006954016433

Torres Labansat, M. (2022). Memorias UNAM 2022. Coordinación General de Estudios de Posgrado. UNAM. Ciudad de Mexico.

Urquhart, S., & Weir, C. (1998). Reading in a Second Language: Process. *Product and*.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2004). Conceptual and methodological issues in adapting tests. En R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 51–76). Psychology Press.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.

Weir, C., Huizhong, Y., & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes* (Número 12). Cambridge University Press.

Weir, C. J. (2003). A survey of the history of the Cerificate of Proficiency in English (CPE) in the twentieth century. En C. J. Weir & M. Milanovic (Eds.), *Continuity and Innovation: The History of the CPE 1913–2002* (pp. 1–56). Cambridge University Press.

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan*, *10*, 9780230514577.

Westchester Institute for Human Services Research. (2003). *High – stakes testing. The Balanced View*. *7*(1).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Worrell, F. C. (2016). Commentary on perspectives on fair assessment. En N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 299–310). Routledge.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187–213.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, *160*.

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

IX. ANEXOS

Anexo 1. Código de R para resultados de examen de comprensión lectora

```
#librerias a utilizar

library(psych)
library(psy)
library(mirt)
library(sjPlot)
library(difR)
library(ltm)
library(lme4)
library(lordif)
library(ShinyItemAnalysis)
library(tidyverse)
library(gt)
library(kableExtra)
library(corrplot)
library(RColorBrewer)
library(eRm)
library(ggdist)



#lectura de las bases de datos en archivo temporal de lista
datatemp <- list.files(pattern="*Lectura")

##lectura de las bases de datos con read_fwf
lectura_list <- lapply(datatemp, read_fwf, col_positions =
                           fwf_widths(c(1,2, 2,5,2,1,1,1,3,1,2,1,1,1,1,1,1,1,1,1,1)
,
                                c("area", "posgrado", "periodo", "ID", "edad",
                                  "genero", "nacionalidad", "nivel", "lic_proced",
                                  "lic_UNAM", "plantel_UNAM", "LR1", "LR2", "LR3",
                                  "LR4", "LR5", "LR6", "LR7", "LR8", "LR9", "LR10",
                                  "LR11")), col_types ="fffciffffffccccccccc")


#Transformación de lista en una sola base de datos por examen
lectura <- as.data.frame(bind_rows(lectura_list))



reactivos_lectura <- select(lectura, LR1:LR11)

#Definición de cadenas de respuesta correcta para cada examen

## Por motivos de confidencialidad se definen todas las respuestas correctas como A
cadena_lectura <- c("A","A","A","A","A","A", "A","A","A","A","A","A")

#Transformaciones de opciones de respuesta a correctas e incorrectas
reactivos_lectura[] <- +t(t(reactivos_lectura) == cadena_lectura)
```

```r
# Base de datos con reactivos convertidos y con demográficos
lectura_dicotomicos <- lectura%>%
  dplyr::select(area, posgrado, edad, genero, nivel)%>%
  bind_cols(reactivos_lectura)

lecturaDIF <- lectura%>%
  dplyr::select(area)%>%
  bind_cols(reactivos_lectura)


## Figura 1 - Distribución de puntajes por área
lectura_dicotomicos%>%
  mutate(Total = rowSums(reactivos_lectura))%>%
  drop_na(genero)%>%
  ggplot(aes(x = area, y = Total, color = area)) +
  ggdist::stat_halfeye(adjust =.8, width = .5, justification = -.5, aes(fill=area)) +
  geom_boxplot(width = .25, outlier.shape = NA) +
  geom_point(size = 1.3, alpha = .4, position = position_jitter(seed =1, width = .4))
+
  coord_cartesian(xlim = c(1.2, NA)) + theme_light() +
  labs(title = "Figura X. Distribución de puntajes en examen de lectura y gramática p
or área de posgrado",
       x = "Área", y = "Puntaje total")+
  scale_fill_manual(values = c("#44BEFE", "#098BD1", "#0D547A", "#092A3C"))+
  scale_color_manual(values = c("#44BEFE", "#098BD1", "#0D547A", "#092A3C"))


lectura_t <- lectura_dicotomicos%>%
  mutate(Total = rowSums(reactivos_lectura))

anova_area<-aov(Total ~ area, data= lectura_t)
summary(anova_area)

TukeyHSD(anova_area)


# Figura 2 -  Porcentaje de respuestas correctas por área

porcentaje <- function(x)
{
  round(sum((x)/n()*100),1)
}

reactivos_lectura%>%
  cbind(lectura$area)%>%
  rename(area = `lectura$area`)%>%
  group_by(area)%>%
  drop_na()%>%
  summarise(across(everything(), list(`%` = porcentaje, n = sum)))%>%
  pivot_longer(!area,
               names_to = c("reactivo",".value"),
               names_pattern = "(.*)_(.)")%>%
  ggplot(aes(x = factor(reactivo, levels = names(reactivos_lectura)), y = `%`, fill =
area)) + geom_col(position = "dodge") +
```

```
  theme_light() +
  labs(title = "Figura X. Porcentaje de respuestas correctas por reactivo en examen d
e lectura y gramática por área de posgrado",
       x = "Reactivo", y = "Porcentaje de respuestas correctas") +
  scale_fill_manual(values = c("#44BEFE", "#098BD1", "#0D547A", "#092A3C"))




#Ejecución de la función par por par
MHRG12<-DIFMH(lecturaDIF, 1,2)
MHRG13<-DIFMH(lecturaDIF, 1,3)
MHRG14<-DIFMH(lecturaDIF, 1,4)
MHRG23<-DIFMH(lecturaDIF, 2,3)
MHRG24<-DIFMH(lecturaDIF, 2,4)
MHRG34<-DIFMH(lecturaDIF, 3,4)

#Generación de la tabla con los resultados de todas las comparativas
TablaDIF_RG_MH<- cbind(names(reactivos_lectura),MHRG12,MHRG13,MHRG14,MHRG23,MHRG24,MH
RG34)

TablaDIF_RG_MH%>%as_tibble%>%gt
# Estilización de la tabla
TablaDIF_RG_MH%>%
  kbl(caption = "Tabla X. Presencia de DIF Área vs Área en Examen de lectura con Méto
do Mantel-Haensze")%>%
  kable_classic(full_width = F, html_font = "Cambria")%>%
  add_header_above(c(" " = 1, "FMI y CBQS" = 2, "FMI y CS" = 2, "FMI y HA" = 2,
                     "CBQS y CS" = 2, "CBQS y HA" = 2, "CS y HA" = 2))


# Análisis con método de regresión logística

#DIF por área con Regresión Logística para examen de lectura
difLogistic(reactivos_lectura, lectura$area, member.type = "cont", purify = TRUE)

#Ejecución de la función par por par - DIF uniforme
LRRG12U<-DIFLR(lecturaDIF, 1,2, "udif")
LRRG13U<-DIFLR(lecturaDIF, 1,3, "udif")
LRRG14U<-DIFLR(lecturaDIF, 1,4, "udif")
LRRG23U<-DIFLR(lecturaDIF, 2,3, "udif")
LRRG24U<-DIFLR(lecturaDIF, 2,4, "udif")
LRRG34U<-DIFLR(lecturaDIF, 3,4, "udif")

#Generación de la tabla con los resultados de todas las comparativas
TablaDIF_RG_LRU<- cbind(names(reactivos_lectura),LRRG12U,LRRG13U,LRRG14U,LRRG23U,LRRG
24U,LRRG34U)

TablaDIF_RG_LRU %>%as_tibble%>%gt
# Estilización de la tabla
TablaDIF_RG_LRU%>%
  as_tibble()%>%
  apa("Tabla X.  Análisis DIF exploratorio del examen de lectura y gramática con
      método recursivo de regresión logística para DIF uniforme")%>%
```

```
  tab_style(style = cell_text(color = "#00B137", weight = "bold"),
            locations = cells_body(columns = 2,rows = LRRG12U > 0.035))%>%
  cols_label(V1 = "Reactivo", LRRG12U = "FMI y CBQS", LRRG13U = "FMI y CS", LRRG14U =
"FMI y HA",
             LRRG23U =  "CBQS y CS", LRRG24U = "CBQS y HA",LRRG34U =  "CS y HA")

#Ejecución de la función par por par - DIF no uniforme
LRRG12NU<-DIFLR(lecturaDIF, 1,2, "nudif")
LRRG13NU<-DIFLR(lecturaDIF, 1,3, "nudif")
LRRG14NU<-DIFLR(lecturaDIF, 1,4, "nudif")
LRRG23NU<-DIFLR(lecturaDIF, 2,3, "nudif")
LRRG24NU<-DIFLR(lecturaDIF, 2,4, "nudif")
LRRG34NU<-DIFLR(lecturaDIF, 3,4, "nudif")

#Generación de la tabla con los resultados de todas las comparativas
TablaDIF_RG_LRU<- cbind(names(reactivos_lectura),LRRG12NU,LRRG13NU,LRRG14NU,LRRG23NU,
LRRG24NU,LRRG34NU)

TablaDIF_RG_LRU %>%as_tibble%>%gt

# Estilización de la tabla
TablaDIF_RG_LRU%>%
  as_tibble()%>%
  apa("Tabla X.  Análisis DIF exploratorio del examen de lectura y gramática con
      método recursivo de regresión logística para DIF no uniforme")%>%
  cols_label(V1 = "Reactivo", LRRG12NU = "FMI y CBQS", LRRG13NU = "FMI y CS", LRRG14N
U = "FMI y HA",
             LRRG23NU =  "CBQS y CS", LRRG24NU = "CBQS y HA",LRRG34NU =  "CS y HA")



## DIF Rasch
#Ejecución de la función par por par - DIF uniforme
RaschCL12<-DIFRasch(lecturaDIF, 1,2)
RaschCL13<-DIFRasch(lecturaDIF, 1,3)
RaschCL14<-DIFRasch(lecturaDIF, 1,4)
RaschCL23<-DIFRasch(lecturaDIF, 2,3)
RaschCL24<-DIFRasch(lecturaDIF, 2,4)
RaschCL34<-DIFRasch(lecturaDIF, 3,4)

TablaDIF_CL_Rasch<- cbind(names(reactivos_lectura),RaschCL12,RaschCL13,RaschCL14,Rasc
hCL23,RaschCL24,RaschCL34)

TablaDIF_CL_Rasch%>%as_tibble%>%gt
## DIF SIBTEST
#Ejecución de la función par por par - DIF uniforme

lecturaDIF1<-lecturaDIF%>%
  select(area,LR1:LR11)%>%
  drop_na()

LORDRG12<-DIF_LORD(lecturaDIF1,1,2)
LORDRG13<-DIF_LORD(lecturaDIF1, 1,3)
LORDRG14<-DIF_LORD(lecturaDIF1, 1,4)
```

```
LORDRG23<-DIF_LORD(lecturaDIF1, 2,3)
LORDRG24<-DIF_LORD(lecturaDIF1, 2,4)
LORDRG34<-DIF_LORD(lecturaDIF1, 3,4)

reactivos_lr <- names(reactivos_lectura)
#Generación de la tabla con los resultados de todas las comparativas
TablaDIF_RG_SIB<- cbind(reactivos_lr,LORDRG12,LORDRG13,LORDRG14,LORDRG23,LORDRG24,LOR
DRG34)

TablaDIF_RG_SIB%>%as_tibble%>%gt
# Estilización de la tabla
TablaDIF_RG_SIB%>%
  as_tibble()%>%
  apa("Tabla X.  Análisis DIF  del examen de lectura y gramática con
      método SIBTEST")%>%
  cols_label(reactivos_rg = "Reactivo", SIBRG12 = "FMI y CBQS", SIBRG13 = "FMI y CS",
SIBRG14 = "FMI y HA",
             SIBRG23 =  "CBQS y CS", SIBRG24 = "CBQS y HA",SIBRG34 =  "CS y HA")

## Regresión logística recursiva



### DIF con 2 grupos. Comparación de métodos
lectura_2grps <- lecturaDIF%>%
  mutate(area = case_when(area %in% 1:2 ~ 1, area %in% 3:4 ~ 2))%>%
  mutate(area = as.factor(area))%>%
  select(area,LR1:LR11)%>%
  drop_na()

LORDRG2gps<-DIF_LORD(lectura_2grps, 2,1)
LRRG2gps<-DIFLR(lectura_2grps, 2,1, "udif")
MHRG2gps<-DIFMH(lectura_2grps, 2,1)
RaschRG2gps<-DIFRasch(lectura_2grps, 2,1)

#Generación de la tabla con los resultados de todas las comparativas
TablaDIF_RG2gps<- cbind(names(reactivos_lectura),MHRG2gps,LRRG2gps,RaschRG2gps,LORDRG
2gps)

TablaDIF_RG2gps%>%
  as_tibble()%>%gt
# Estilización de la tabla
TablaDIF_RG2gps%>%
  as_tibble()%>%
  apa("Tabla X.  Análisis DIF  del examen de lectura y gramática c- Comparación de 2
grupos con métodos tradicionales")%>%
  cols_label(V1 = "Reactivo", SIBRG2gps = "SIBTEST", LRRG2gps = "Regresión Logística"
, MHRG2gps = "Mantel-Haenszel" ,
             RaschRG2gps =  "Diferencia de Logits (Rasch)")


## DIFtree
library(DIFtree)
```

```r
covar_lectura <- lectura%>%
  select(!edad)%>%
  select(area,posgrado, genero, nivel)%>%
  mutate_all(as.numeric)


diftree <- reactivos_lectura%>%
  drop_na()%>%
  as.matrix()



nudiftree_lectura<- DIFtree(Y = diftree, X = covar_lectura,
                            model = "Logistic", type = "nudif")

tabla_logtree <- summary(nudiftree_lectura)

tabla_logtree[[2]]%>%
  dplyr::rename(Tipo = type, Particiones = nosplits)%>%
  apa("Tabla 9.  Análisis DIF exploratorio del examen de lectura y gramática con
      método recursivo de regresión logística")



## Definición de función para obtener valores de DeltaMH y significancia con base en
función RM y Waldtest de paquete eRm
library(eRm)
DIFRasch <- function(base, referencia, focal){

  areas <- base%>%
    filter(area %in% c(referencia,focal))

  reactivos <- areas%>%
    select(!area:plantel_UNAM)

  modeloR<-RM(reactivos)
  subgroup_diffs <- Waldtest(modeloR, splitcr = areas$area)
  subgroup_1_diffs <- subgroup_diffs$betapar1
  subgroup_2_diffs <- subgroup_diffs$betapar2
  coef <- round((subgroup_1_diffs*-1)-(subgroup_2_diffs*-1),3)

  DIF <- symnum(coef, c(-Inf,-1,-.5, .5, 1, Inf), symbols = c("--","-","", "+", "++")
)

  tabla_dif<-cbind(coef, DIF)
}

#Ejecución de la función par por par

RaschRG12<-DIFRasch(lecturaDIF, 1,2)
RaschRG13<-DIFRasch(lecturaDIF, 1,3)
RaschRG14<-DIFRasch(lecturaDIF, 1,4)
```

```
RaschRG23<-DIFRasch(lecturaDIF, 2,3)
RaschRG24<-DIFRasch(lecturaDIF, 2,4)
RaschRG34<-DIFRasch(lecturaDIF, 3,4)
RaschRG12<-DIFRasch(lecturaDIF, 1,2)
RaschRG13<-DIFRasch(lecturaDIF, 1,3)
RaschRG14<-DIFRasch(lecturaDIF, 1,4)
RaschRG23<-DIFRasch(lecturaDIF, 2,3)
RaschRG24<-DIFRasch(lecturaDIF, 2,4)
RaschRG34<-DIFRasch(lecturaDIF, 3,4)

#Generación de la tabla con los resultados de todas las comparativas

TablaDIF_RG_Rasch<- cbind(names(reactivos_lectura),RaschRG12,RaschRG13,RaschRG14,Rasc
hRG23,RaschRG24,RaschRG34)

# Estilización de la tabla
TablaDIF_RG_Rasch%>%
  kbl(caption = "Tabla 8. Presencia de DIF Área vs Área en Examen de lectura con Méto
do de Rasch")%>%
  kable_classic(full_width = F, html_font = "Cambria")%>%
  add_header_above(c(" " = 2, "FMI y CBQS" = 2, "FMI y CS" = 2, "FMI y HA" = 2,
                     "CBQS y CS" = 2, "CBQS y HA" = 2, "CS y HA" = 2))



### Análisis DIF con dos grupos I y II vs III HA

#Comparativa de distintos modelos DIF en examen de lectura por género
dif2gps_lectura <-  dichoDif(lectura_2grps,group=12, focal.name=1,
                             method=c("MH","Logistic","Lord","SIBTEST"),
                             model="2PL")


difMH(lectura_2grps,group=12, focal.name=1)

difLogistic(lectura_2grps,group=12, focal.name=1)


difLord(lectura_2grps,group=12, focal.name=1,model="2PL")


difSIBTEST(lectura_2grps,group=12, focal.name=1)%>%
  str()

## Análisis de cumplimiento de supuestos IRT
#Supuestos IRT
install.packages("ltm")
install.packages("TAM")
library(ltm)
library(TAM)

# Comparación de modelos IRT
RG1PL <- ltm::rasch(reactivos_lectura)
```

```
RG2PL <- ltm(reactivos_lectura ~ z1)
RG3PL <- tpm(reactivos_lectura)

anova(RG2PL, RG3PL)

#Supuesto de independencia local
TAMRG2PL <- tam.mml.2pl(reactivos_lectura, irtmodel = "2PL")
tamRGfit <- TAM::tam.modelfit(TAMRG2PL)
RG_Q3_items <- tamRGfit$stat.itempair
round(RG_Q3_items[,5:8],3)
-1/(nrow(reactivos_lectura)-1)

corrplot(tamRGfit$Q3.matr,  method = "number", type="lower", col=brewer.pal(n=11, nam
e="RdBu"),tl.srt=90)


residuals(RG_mirt2, type = "Q3")%>%
  corrplot(method = "number", type="lower", col=brewer.pal(n=11, name="RdBu"),tl.srt=
90)


residuals(RG_mirt1)
# Figura 3 - Mapa de Wright


## Mapa de Wright con modelo Rasch
library(ShinyItemAnalysis)
library(WrightMap)
library(RColorBrewer)
modeloRasch_RG <- mirt(data=reactivos_lectura,model = 1, itemtype="Rasch",
                       SE=TRUE, verbose=FALSE)

thetaRaschRG1 <- fscores(modeloRasch_RG)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 1)%>%
  select(F1)

thetaRaschRG2 <- fscores(modeloRasch_RG)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 2)%>%
  select(F1)

thetaRaschRG3 <- fscores(modeloRasch_RG)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 3)%>%
  select(F1)

thetaRaschRG4 <- fscores(modeloRasch_RG)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 4)%>%
```

```
  select(F1)


coef.Rasch_RG <- coef(modeloRasch_RG, IRTpars=TRUE, simplify=TRUE)

b_RaschRG <- coef.Rasch_RG[[1]][,2]

item_color_palette <- matrix(rep(c("#44BEFE","#098BD1","#0D547A","#092A3C")),
                             10, byrow = TRUE, ncol = 4)

split.screen(figs = matrix(c(    0, .10,0,1,
                                 .10, .20,0,1,
                                 .20, .30,0,1,
                                 .30, .40,0,1,
                                 .40,  1, 0,1), ncol = 4, byrow = TRUE))
personHist(thetaRaschRG1,yRange = c(-1.2,1.5), dim.lab.cex = 1,
           dim.names = "Área 1", dim.color = "#44BEFE",
           show.axis.logits = FALSE, axis.persons = "Sustentantes", breaks = 10)

screen(2)
personHist(thetaRaschRG2,yRange = c(-1.2,1.5), dim.lab.cex = 1,
           dim.names = "Área 2", dim.color = "#098BD1",
           show.axis.logits = FALSE, axis.persons = NULL, breaks = 10)

screen(3)
personHist(thetaRaschRG3,yRange = c(-1.2,1.5), dim.lab.cex = 1,
           dim.names = "Área 3", dim.color = "#0D547A",
           show.axis.logits = FALSE, axis.persons = NULL, breaks = 10)

screen(4)
personHist(thetaRaschRG4, yRange = c(-1.2,1.5), dim.lab.cex = 1,
           dim.names = "Área 4", dim.color = "#092A3C",
           show.axis.logits = FALSE,axis.persons = NULL, breaks = 10)

screen(5)
itemModern(b_RaschRG, yRange = c(-1.2,1.5), thr.sym.cex = 2, thr.sym.col.bg = "#092A3
C")
mtext("Figura 3. Mapa de Wright del exámen de lectura y gramática con distribución po
r área", side = 3, font = 2, line = 1)

close.screen(all.screens = TRUE)

# Figura 4 - Mapa de Wright con modelo de Rasch


## Mapa de Wright con modelo Rasch y cuatro áreas
modeloRasch_RG_A <- multipleGroup(data=reactivos_lectura,model = 1, group = lectura$a
rea,
                                  itemtype="Rasch", SE=TRUE)

thetaRaschRG1A <- fscores(modeloRasch_RG_A)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
```

```
  filter(area == 1)%>%
  select(F1)

thetaRaschRG2A <- fscores(modeloRasch_RG_A)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 2)%>%
  select(F1)

thetaRaschRG3A <- fscores(modeloRasch_RG_A)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 3)%>%
  select(F1)

thetaRaschRG4A <- fscores(modeloRasch_RG_A)%>%
  cbind(area = lectura$area)%>%
  as_tibble()%>%
  filter(area == 4)%>%
  select(F1)


coef.Rasch_RG_A <- coef(modeloRasch_RG_A, IRTpars=TRUE, simplify=TRUE)
b_RaschRG1 <- coef.Rasch_RG_A[[1]][[1]][,2]
b_RaschRG2 <- coef.Rasch_RG_A[[2]][[1]][,2]
b_RaschRG3 <- coef.Rasch_RG_A[[3]][[1]][,2]
b_RaschRG4 <- coef.Rasch_RG_A[[4]][[1]][,2]

b_RaschRGA <- cbind(b_RaschRG1, b_RaschRG2, b_RaschRG3, b_RaschRG4)

item_color_palette <- matrix(rep(c("#44BEFE","#098BD1","#0D547A","#092A3C")),
                             11, byrow = TRUE, ncol = 4)

split.screen(figs = matrix(c(    0, .10,0,1,
                                 .10, .20,0,1,
                                 .20, .30,0,1,
                                 .30, .40,0,1,
                                 .40,  1, 0,1), ncol = 4, byrow = TRUE))
personHist(thetaRaschRG1A,yRange = c(-1.5,1), dim.lab.cex = 1,
           dim.names = "Área 1", dim.color = "#44BEFE",
           show.axis.logits = FALSE, axis.persons = "Sustentantes", breaks = 10)

screen(2)
personHist(thetaRaschRG2A,yRange = c(-1.5,1.5), dim.lab.cex = 1,
           dim.names = "Área 2", dim.color = "#098BD1",
           show.axis.logits = FALSE, axis.persons = NULL, breaks = 10)

screen(3)
personHist(thetaRaschRG3A,yRange = c(-1.5,1.5), dim.lab.cex = 1,
           dim.names = "Área 3", dim.color = "#0D547A",
           show.axis.logits = FALSE, axis.persons = NULL, breaks = 10)

screen(4)
personHist(thetaRaschRG4A, yRange = c(-1.5,1.5), dim.lab.cex = 1,
```

```
           dim.names = "Área 4", dim.color = "#092A3C",
           show.axis.logits = FALSE,axis.persons = NULL, breaks = 10)

screen(5)
itemModern(b_RaschRGA, yRange = c(-1.5,1.5), thr.sym.cex = 2, thr.sym.col.bg =
             item_color_palette)
mtext("Figura 4. Mapa de Wright del exámen de lectura y gramática con dificultad calc
ulada por área", side = 3, font = 2, line = 1)

close.screen(all.screens = TRUE)

# Figura 5 - Análisis Raschtree

library(psychotree)
lectura_dicotomicos$resp <- as.matrix(lectura_dicotomicos[, 6:16])

#Eliminr variables no requeridas
lecturaDIF1 <- lectura_dicotomicos[, -(6:16)]

# Creación de modelo psychotree lectura
raschtreeRG1 <- psychotree::raschtree(resp ~ area  + genero + nivel, data = lecturaDI
F1)

#Gráfica del modelo
plot(raschtreeRG1, title = "Figura 5")


# Selección de modelo TRI

CL_irt1<-mirt(reactivos_lectura, model = 1, itemtype = "Rasch")
CL_irt2<-mirt(reactivos_lectura, model = 1, itemtype = "2PL")
CL_irt3<-mirt(reactivos_lectura, model = 1, itemtype = "3PL")

CL_mirt2<-mirt(reactivos_lectura, model = 2, itemtype = "2PL")



residuals(RG_mirt2, type = "LDG2")

residuals(RG_mirt2, type = "Q3")%>%
  corrplot(method = "number", type="lower", col=brewer.pal(n=11, name="RdBu"),tl.srt=
90)




fit_RG_irt1 <- M2(RG_irt1, na.rm = TRUE)
fit_RG_irt2 <- M2(RG_irt2, na.rm = TRUE)
fit_RG_irt3 <- M2(RG_irt3, na.rm = TRUE)


nombres_modelos1 <- c("Modelo Rasch", "Modelo 2PL", "Modelo 3PL")
```

```
rbind(fit_RG_irt1, fit_RG_irt2, fit_RG_irt3)%>%
  select(M2:RMSEA, CFI)%>%
  cbind(nombres_modelos1, .)%>%
  gt("Tabla 3. Comparación de modelos TRI unidimensionales en examen de lectura y gra
mática",rowname_col = "nombres_modelos1")%>%
  tab_stubhead(label = "Modelos")%>%
  fmt_number(columns = c(2,4:6),
             decimals = 3)


anova_irt1_vs_mirt2 <-anova(CL_irt2, CL_mirt2)

Modelos <- c("Modelo 1", "Modelo 2")
anova_irt1_vs_mirt2%>%
  select(AIC, BIC, logLik, X2, df, p)%>%
  cbind(Modelos, .)%>%
  gt("Tabla 4. Comparativa de Modelo unidimensional y multidimensional")%>%
  fmt_number(columns = 2:5,
             decimals = 2)%>%
  tab_footnote(footnote = "Modelo unidimensional",
               locations = cells_body(columns = Modelos, rows = 1))%>%
  tab_footnote(footnote = "Modelo exploratorio de dos dimensiones",
               locations = cells_body(columns = Modelos, rows = 2))



#Generación de modelo de dos parámetros específicos para cada grupo usando multipleGr
oup del paquete mirt
modelo2PL_CL_A <- multipleGroup(data=reactivos_lectura,model = 1, group = lectura$are
a,
                                itemtype="2PL", SE=TRUE)

#Análisis DIF a partir del modelo de dos parámetros anterior
dif_CL1 <-DIF(modelo2PL_CL_A, which.par = c("a1", "d"))
dif_CL1%>%arrange(X2)

# A partir del análisis anterior, se identifican reactivos anRGa y se vuelve a ejecut
ar el análisis fijando los parámetros en dichos reactivos

modelo2PL_CL_ANCH <- multipleGroup(data=reactivos_lectura,model = 1, group = lectura$
area,
                                   itemtype="2PL", SE=TRUE, invariance = c(c("LR2"),
                                                                           'free_mean
s', 'free_var'))

#Análisis DIF ya con los reactivos ancla

  dif_CL2<-DIF(modelo2PL_CL_ANCH, c("a1", "d"), items2test = c(1,3:11), maxiter = 100
0)

      library(difR)
library(mirt)
dif_CL3 <- DIF(modelo2PL_CL_ANCH, which.par = c("a1", "d"), items2test = c(1,3:11), c
ontrol = list(optimizer="quasinewton"), maxiter = 1000)
```

```r
nombres_reactivos2 <- rownames(dif_CL2)

dif_CL2%>%
  cbind(nombres_reactivos2,.)%>%
  rename(Reactivos =nombres_reactivos2)%>%
  gt("Tabla 5. Análisis DIF con método de Razón de Verosimilitud en examen de lectura
y gramática")%>%
  fmt_number(columns = c(3:7,9),
             decimals = 3)

### Comparación DIF con dos grupos MIRT
lectura_2grps <- lecturaDIF%>%
  mutate(area = case_when(area %in% 1:2 ~ 1, area %in% 3:4 ~ 2))%>%
  mutate(area = as.factor(area))%>%
  select(area,LR1:LR11)

modelo2PL_RG2gp <- multipleGroup(data=lectura_2grps[,2:12],model = 1,
                                 group = lectura_2grps$area,
                                 itemtype="2PL", SE=TRUE)

dif_RG2gp<-DIF(modelo2PL_RG2gp, c("a1", "d"), plotdif = TRUE)
dif_RG2gp%>%arrange(X2)



modelo2PL_RG2gp_ANCH <- multipleGroup(data=reactivos_lectura,model = 1,
                                      group = lectura_2grps$area, itemtype="2PL",
                                      SE=TRUE, invariance = c(c("LR2", "LR8", "LR1"),
                                                              'free_means', 'free_var
'))

dif_RG2gp_2<-DIF(modelo2PL_RG2gp_ANCH, c("a1", "d"), items2test = c(3:7,9:11),plotdif
= TRUE)

rownames(dif_RG2gp_2)


plot(modelo2PL_RG2gp_ANCH, type ="trace")
plot(modelo2PL_RG_ANCH, type ="trace")


dif_RG2gp_2%>%
  cbind(rownames(dif_RG2gp_2),.)%>%
  rename(Reactivos = `rownames(dif_RG2gp_2)`)%>%
  apa("Tabla X. Análisis DIF con método de Razón de Verosimilitud en examen de lectur
a y gramática - Comparación con 2 grupos")%>%
  fmt_number(columns = c(3:7,9),
             decimals = 3)
```

Anexo 2. Código de R para resultados de examen de redacción y gramática

```
# Todos los análisis previos a esta sección se realizaron utilizando el mismo código
que en el examen de comprensión lectora, modificando levemente la sintaxis para
adaptarse a la cantidad de reactivos. Por ello, se omite la sintáxis de estos análisi
s para evitar redundancias y se incluye solamente la sección novedosa referente al
modelado de DIF usando modelos TRI más complejos


## Análisis DIF TRI robustos

## Selección de modelo TRI

## Generación de cada modelo

rg_irt1<-mirt(reactivos_redaccion, model = 1, itemtype = "Rasch")
rg_irt2<-mirt(reactivos_redaccion, model = 1, itemtype = "2PL")
rg_irt3<-mirt(reactivos_redaccion, model = 1, itemtype = "3PL")
rg_irt4<-mirt(reactivos_redaccion, model = 1, itemtype = "4PL", method = "QMCEM")


fit_rg_irt1 <- M2(rg_irt1, na.rm=TRUE)
fit_rg_irt2 <- M2(rg_irt2, na.rm=TRUE)
fit_rg_irt3 <- M2(rg_irt3, na.rm=TRUE)
fit_rg_irt4 <- M2(rg_irt4, na.rm=TRUE)

anova(rg_irt3, rg_irt4)

rg_irt1<-mirt(reactivos_redaccion, model = 1, itemtype = "2PL", method = "EM", SE=TRU
E)
rg_mirt2<-mirt(reactivos_redaccion, model = 2, itemtype = "2PL", method = "EM", SE=TR
UE)
rg_mirt3<-mirt(reactivos_redaccion, model = 3, itemtype = "2PL", method = "QMCEM")
rg_mirt4<-mirt(reactivos_redaccion, model = 4, itemtype = "2PL", method = "QMCEM")
rg_mirtc<-mirt(reactivos_redaccion, model = modelo1, itemtype = "2PL", method = "QMCE
M", SE=TRUE)
rg_mirtcpc<-mirt(reactivos_redaccion, model = modelo1, itemtype = "PC2PL", method = "
QMCEM", SE=TRUE)


fit_rg_irt1 <- M2(rg_irt1)
fit_rg_mirt2 <- M2(rg_mirt2)
fit_rg_mirt3 <- M2(rg_mirt3, na.rm = T)
fit_rg_mirt4 <- M2(rg_mirt4, na.rm = T, QMC = T)
fit_rg_mirtc <- M2(rg_mirtc, na.rm = T, QMC = T)
fit_rg_mirtcpc <- M2(rg_mirtcpc, na.rm = T, QMC = T)

nombres_modelos1 <- c("Modelo unidimensional", "Modelo 2 dimensiones",
                      "Modelo 3 dimensiones", "Modelo 4 dimensiones",
                      "Modelo confirmatorio 4 dimensiones",
                      "Modelo confirmatorio parcialmente compensatorio")

rbind(fit_rg_irt1, fit_rg_mirt2, fit_rg_mirt3, fit_rg_mirt4, fit_rg_mirtc, fit_rg_mir
```

```
tcpc)%>%
  select(M2:RMSEA, CFI)%>%
  cbind(nombres_modelos1, .)%>%
  apa("Tabla 12. Comparación de modelos TRI multidimensionales en examen de Redacción
y Gramática",rowname_col = "nombres_modelos1")%>%
  tab_stubhead(label = "Modelos")%>%
  fmt_number(columns = c(2,4:6),
             decimals = 3)



rg_mirt2%>%
  extract.mirt("F")%>%
  as_tibble()%>%
  mutate(F1 = case_when(F1 < -.25~F1), F2 = case_when(F2 < -.25~F2))




reactivos <- names(reactivos_redaccion)
extract.mirt(rg_mirt2, "F")%>%
  as_tibble(.)%>%
  cbind(reactivos,.)%>%
  apa("Tabla X. Cargas de los reactivos de examen de redacción en el modelo explorato
rio de 2 dimensiones")%>%
  fmt_number(columns =c(2:3),
             decimals = 3)%>%
  tab_style(style = cell_text(color = "#00B137", weight = "bold"),
            locations = cells_body(columns = 2,rows = F1 > 0.3 |F1< -0.3))%>%
  tab_style(style = cell_text(color = "#00B137", weight = "bold"),
            locations = cells_body(columns = 3,rows = F2 > 0.3 |F2< -0.3))%>%
  tab_footnote(footnote = "Se señalan en verde los valores con una carga mayor a .3",
               locations = cells_body(columns = F1, rows = 1))


residuals(rg_mirt2, type = "Q3")%>%
  corrplot(method = "number", type="lower", col=brewer.pal(n=11, name="RdBu"),tl.srt=
90)

## Generación de modelo de dos parámetros específicos para cada grupo usando multiple
Group del paquete mirt
### Análisis DIF con  mirt
modelo2PL1_RG_A <- multipleGroup(data=reactivos_redaccion,model = 1, group = redaccio
n$area,
                                 itemtype="2PL", SE=TRUE)




#Análisis DIF a partir del modelo de dos parámetros anterior
cif_RG1<-DIF(modelo2PL1_RG_A, c("a1", "d"),plotdif = TRUE)
dif_RG1%>%arrange(X2)
```

```
## A partir del análisis anterior, se identifican reactivos anclaa y se vuelve a ejec
utar el análisis fijando los parámetros en dichos reactivos
modelo2PL1_RG_ANCH <- multipleGroup(data=reactivos_redaccion,model = 1, group = redac
cion$area,
                                    itemtype="2PL", SE=TRUE, invariance = c(c("RG3","
RG11", "RG15","RG7","RG16"),
                                                                            'free_mea
ns', 'free_var'))

#Análisis DIF ya con los reactivos ancla

dif_RG2<-DIF(modelo2PL1_RG_ANCH, c("a1", "d"), items2test = c(1,2,4:6,8:10,12:14,17:2
0), plotdif = TRUE)

dif_RG2%>%arrange(X2)


## Generación de modelo de dos parámetros específicos para cada grupo usando multiple
Group del paquete mirt
### Análisis DIF con  mirt y modelo bidimensional
modelo2PL2_RG_A <- multipleGroup(data=reactivos_redaccion,model = 2, group = redaccio
n$area,
                                 itemtype="2PL", SE=TRUE)




#Análisis DIF a partir del modelo de dos parámetros anterior
dif_RG1<-DIF(modelo2PL2_RG_A, c("a1", "d"),plotdif = TRUE)
dif_RG1%>%arrange(X2)

## A partir del análisis anterior, se identifican reactivos anclaa y se vuelve a ejec
utar el análisis fijando los parámetros en dichos reactivos
modelo2PL2_RG_ANCH <- multipleGroup(data=reactivos_redaccion,model = 2, group = redac
cion$area,
                                    itemtype="2PL", SE=TRUE, invariance = c(c("RG3","
RG11", "RG15","RG7","RG16"),
                                                                            'free_mea
ns', 'free_var'))

#Análisis DIF ya con los reactivos ancla

dif_RG2<-DIF(modelo2PL2_RG_ANCH, c("a1", "d"), items2test = c(1,2,4:6,8:10,12:14,17:2
0), plotdif = TRUE)

dif_RG2%>%arrange(X2)


## generación de la tabla final

dif_RG2%>%
  cbind(nombres_reactivos2,.)%>%
  rename(Reactivos =nombres_reactivos2)%>%
```

```
  gt("Análisis DIF con método de Razón de Verosimilitud en examen de redacción y gram
ática")%>%
  fmt_number(columns = c(3:7,9),
             decimals = 3)

### Comparación DIF con dos grupos MIRT
redaccion_2grps <- redaccion%>%
  mutate(areas = case_when(area %in% 1:2 ~ 1, area %in% 3:4 ~ 2))%>%
  mutate(areas = as.factor(areas))

modelo2PL_rg2gp <- multipleGroup(data=reactivos_redaccion,model = 2,
                                 group = redaccion_2grps$areas,
                                 itemtype="2PL", SE=TRUE)

dif_rg2gp<-DIF(modelo2PL_rg2gp, c("a1", "d"))
dif_rg2gp%>%arrange(X2)


## Modelo con reactivos ancla

modelo2PL_rg2gp_ANCH <- multipleGroup(data=reactivos_redaccion,model = 2,
                                      group = redaccion_2grps$areas, itemtype="2PL",
                                      SE=TRUE, invariance = c(c("RG15", "RG8", "RG9",
                                                                "RG7", "RG12", "RG16"
,
                                                                "RG3", "RG10", "RG5",
                                                                "RG13", "RG2", "RG11"
),
                                                               'free_means', 'free_var
'))

dif_RG2gp_2<-DIF(modelo2PL_rg2gp_ANCH, c("a1", "d"), items2test = c(1,4,6,14,17
                                                                    ,18,19,20))
reactivosRG2gp <- rownames(dif_RG2gp_2)

## Gráficos de CCI complementarios
plot(modelo2PL_RG2gp_ANCH, type ="trace")
plot(modelo2PL_RG_ANCH, type ="trace")


## Tabla final

dif_RG2gp_2%>%
  cbind(rownames(dif_RG2gp_2),.)%>%
  rename(Reactivos = `rownames(dif_RG2gp_2)`)%>%
  apa("Tabla X. Análisis DIF con método de Razón de Verosimilitud en examen de redacc
ión y gramática - Comparación con 2 grupos")%>%
  fmt_number(columns = c(3:7,9),
             decimals = 3)
```