UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

IMPACTO DEL USO DE ESTRATEGIAS DE RESOLUCIÓN BASADAS EN LA POSICIÓN DE LAS OPCIONES EN LOS PUNTAJES OBTENIDOS EN PRUEBAS DE SELECCIÓN MÚLTIPLE: UN ESTUDIO DE SIMULACIONES

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO

GABRIEL FABIÁN ORTEGA HERNÁNDEZ

PROFESOR GUÍA:
SÉVERIN LIONS

PROFESOR CO-GUÍA:
PABLO DARTNELL ROY

COMISIÓN:
JAIME SAN MARTIN ARISTEGUI

Este trabajo ha sido parcialmente financiado por por los proyectos FONDEF ID16I20090, ANID/PIA/Fondos Basales para Centros de Excelencia FB0003, y CMM ANID BASAL FB210005

SANTIAGO DE CHILE
2023

Resumen

La distribución de las opciones correctas en las pruebas de selección múltiple ha mostrado ser una variable psicométrica relevante, en parte debido a que estrategias de resolución basadas en la posición de las opciones permiten potencialmente sacar provecho de ella y, por tanto, amenazan la validez de los resultados. Distintas recomendaciones para el posicionamiento de las opciones correctas de una prueba han sido proporcionadas a los constructores para evitar entregar pistas a los examinados. Aunque aún no existe un consenso entre autores en estas recomendaciones, entre ellas destacan aleatorizar, balancear globalmente y/o balancear localmente las claves en la prueba. También han sido documentadas estrategias de resolución basadas en la posición de las opciones que podrían permitir sacar ventaja de la distribución de las opciones correctas. Sin embargo, no hay claridad acerca de cuanta ventaja en los puntajes obtenidos en pruebas puede otorgar cada una de estas estrategias frente a cada método de posicionamiento y cuál sería entonces la mejor recomendación.

En este trabajo, distintas recomendaciones sobre posicionamiento de opciones correctas en pruebas han sido modeladas, y la efectividad del uso de estrategias basadas en la posición de las opciones ha sido analizada mediante un estudio de simulaciones. Se ha considerado para ello pruebas con distinto número de ítems, con distinto número de opciones, y examinados con distintos niveles de desempeño.

Se encuentra que no preocuparse por la posición de las opciones correctas en una prueba de selección múltiple pone en riesgo la validez de los resultados de la prueba. Similarmente, cuando se siguen recomendaciones de balanceo, sea global o local, se observa que surgen problemas de validez, a los cuales se agregan problemas de inequidad en la ventaja que se obtiene del uso de estrategias. Estos resultados se han verificado para pruebas con distinto número de ítems, distinto número de opciones, y examinados con distinto nivel de desempeño.

Se concluye que randomizar la posición de la opción correcta es la única recomendación que no genera problemas de validez o equidad en los resultados frente al uso de estrategias basadas en la posición de las opciones, sean la que sean. Todos los otros modelos mostraron permitir ventajas significativas con respecto al uso de estrategias basadas en la posición de las opciones. Randomizar aparece como el mejor método de posicionamiento de opciones, y eso independientemente del número de ítems, número de opciones y nivel de desempeño de los examinados.

Agradecimientos

En primer lugar quiero agradecer a mi familia, a mi padre José y a mi madre Georgina por enseñarme todo lo que sé, y por ser un apoyo incondicional mi vida entera. A mi hermana por su cariño y la infinita paciencia que me ha tenido por años y años. Especialmente quiero agradecer a mi abuela Nilda, por ser un ejemplo a seguir para mi, y por enseñarme el valor del trabajo duro y por convencerme de que uno puede imponerse al destino.

Muchas gracias a Severin por su incansable apoyo a lo largo de esta tesis, y durante todos los años que hemos trabajado juntos. Gracias por enseñarme el rigor de trabajar en ciencias y por ayudarme a afrontar los problemas de la vida misma, ha sido una experiencia invaluable haberlo conocido. También quiero agradecer al profesor Pablo Dartnell, junto con Severin me abrieron las puertas para trabajar en educación, y han sido un apoyo constante en mi vida laboral en el CIAE. Quiero agradecer También al profesor Jaime San Martin por sus consejos y ayuda a lo largo de toda la carrera.

Agradezco también la ayuda que he recibido de los funcionarios del DIM, y especialmente a doña Gladys, don Oscar y doña Maria Cecilia que siempre mostraron preocupación por mi persona.

Quiero agradecer también a mis amigos del departamento, que se han convertido en pilares fundamentales para poder dedicarme a la matemática: Victor, Juanco, Toby, Manu y Flores. Les estaré siempre agradecido por pasarlo tan bien junto a ustedes dentro y fuera de la matemática. También quiero agradecer a mis amigos de toda la vida: Francisco, Fabián y Gabriel Morales. Su cariño y compañia a lo largo de mi vida ha sido muy importante para ser la persona que soy. Por último, quiero agradecer a mi eterno amigo Marcos Maldonado, su hermosa amistad siempre estará en mi corazón y esta tesis está dedicada especialmente a él, estoy seguro que se alegraría mucho, más siempre quiso lo mejor para mi.

Tabla de Contenido

Capítulo 1

Introducción

La posición de las opciones correctas es un aspecto a tomar en cuenta en la confección de pruebas de selección múltiple. Autores de guías para la construcción de ítems y pruebas han proporcionado diferentes recomendaciones para el posicionamiento de la alternativa correcta (e.g., [14, 16, 17]). Si bien no existe consenso todavía, de estas guías se destacan tres recomendaciones principales: Randomizar la posición de la opción correcta para cada ítem, balancear esta posición al nivel de la prueba (i.e., colocar la opción correcta en las posibles posiciones el mismo número de veces), y balancear a nivel local la posición (i.e., evitar que una misma posición aloje la opción correcta repetitivamente). Estas recomendaciones buscan que los examinados no puedan identificar la opción correcta en base a su posición [24].

A pesar de la existencia de estas recomendaciones, todavía se observan sesgos de posición con respecto al posicionamiento de la alternativa correcta por parte de los constructores de pruebas. El fenómeno más destacado es la tendencia de los constructores a colocar más frecuentemente la opción correcta en una posición céntrica, fenómeno llamado "middle bias" [1]. El sesgo hacia el medio de los constructores se ha observado en números estudios analizando ítems o pruebas ensambladas de 3, 4 y 5 opciones [25].

El principal problema subyacente a la existencia de sesgos en el posicionamiento de las claves es que a veces los examinados identifican exitosamente la opción correcta a un ítem gracias al uso de estrategias de resolución basadas en la posición de las opciones [7]. Los puntos extras que puedan obtener mediante esta conducta estratégica inyectan ruido en la medición, representando una amenaza para la validez de los resultados de pruebas [15]. Existen diferentes estrategias de resolución basadas en la posición de las opciones; algunas buscan sacar provecho de la presencia potencial del sesgo hacia el medio; otras, pretenden adaptar la distribución de respuestas a la distribución de opciones correctas elegida por el constructor. Las principales estrategias que se han documentado son: Escoger la opción con menor frecuencia en las respuestas para aprovechar el balanceo global de una prueba [3], evitar seleccionar la misma opción para responder a ítems succesivos [23], y escoger una sola posición para rellenar las respuestas faltantes, siguiendo la intuición popular "when in doubt, choose C" [8].

En este estudio, se propone estudiar el impacto que puede tener el uso de cada una de estas estrategias en el desempeño de los examinados. La ventaja asociada a cada estrategia

se compara con los puntos extras obtenidos mediante el uso de una estrategia control, que consiste en elegir aleatoriamente (de manera uniforme) la posición de las respuestas a ítems no respondidos por conocimiento. Se estudia el impacto de estas estrategias en el caso de rendir distintos tipos de pruebas: Con una distribución de alternativas correctas balanceada globalmente y/o localmente entre posiciones, sesgada hacia el centro, o con un posicionamiento aleatorio de las opciones correctas. Se analiza además la interacción que tiene este uso de estrategias con el rendimiento académico de los examinados y con el número de ítems de la prueba. Finalmente, todos los análisis se realizan primero para pruebas de 4 opciones, número más frecuente de opciones en bancos de ítems internacionales [25], pero se replican luego para pruebas de 3 y pruebas de 5 opciones.

El objetivo principal del estudio es doble: Primero, medir la magnitud del ruido en la medición (con respecto a responder aleatoriamente) inyectada por el uso de las estrategias de resolución basadas en la posición de opciones en diferentes tipos de pruebas y para diferentes estudiantes; y segundo, cuestionar si los constructores deben o no preocuparse por su estrategia de posicionamiento de opciones e identificar la potencial mejor estrategia. Gracias al uso de simulaciones matemáticas, se deberían poder identificar las prácticas de constructores que más aseguran la validez de los resultados, y proporcionar directrices sobre la construccción de pruebas que sean basadas en la evidencia, respondiendo así a un llamado recurrente de expertos del campo [16, 17, 18].

Considerando que el formato de selección múltiple es probablemente el que más se usa en el mundo, en pruebas aplicadas en aula por profesores y en pruebas estandarizadas, y que se usa en las pruebas de acceso a la educación superior en Chile, los resultados de este estudio tienen el potencial de interesar un amplio público en el ámbito de la educación, sea nacional o internacional.

Capítulo 2

Estudio 1: Ganancia del uso de estrategias en pruebas de 4 opciones

Este estudio evaluó el impacto de estrategias de resolución basadas en la posición de las opciones, que estaban documentadas en la literatura, en un contexto de pruebas de selección múltiple construidas en base a algunos protocolos reportados para el posicionamiento de opciones correctas (también llamadas claves).

Fueron consideradas pruebas de 4 opciones, porque este número de opciones fue encontrado ser usado más frecuentemente [25]. Por este motivo, los Estudios 2 y 3 también consideraron pruebas de 4 opciones.

2.1. Método

En la sección diseño, los principales objetos de estudio son definidos matemáticamente y explicados. También es explicado el esquema de simulación, y los indicadores y parámetros que serán estudiados y analizados. En la sección modelos de contrucción, los modelos de construcción son definidos y también se justifica la elección de sus principales parámetros. En la sección de estrategias de resolución, las estrategias son definidas operacionalmente. Por último, los parámetros a analizar se especifican en la sección de análisis.

2.1.1. Diseño

Para estudiar la ganancia en puntos extra de una estrategia de resolución basada en la posición de las opciones se modeló el proceso de resolución de una prueba en dos etapas: El examinado responde primero un conjunto de ítems de la prueba en base a su conocimiento curricular, y después responde los ítems faltantes usando una estrategia basada en la posición de las anteriores respuestas por conocimiento.

Como se quería estudiar distintas formas de posicionar las claves, una prueba fue modelada

como una secuencia ordenada donde cada elemento de la secuencia representó la posición de la clave de un ítem. Fijando parámetros, una prueba de N ítems de selección múltiple con un número $N_{opciones}$ de opciones, es un elemento p del conjunto de secuencias $\mathcal{A}^N$, con $\mathcal{A}$ el conjunto de las opciones (en el caso de 4 opciones $\mathcal{A}$ sería $\{a, b, c, d\}$). Para modelar la forma en que las claves son posicionadas por el constructor, se consideró $\mathcal{A}^N$ equipado con una medida de probabilidad $\mathbb{P}$ dada por un modelo de construcción de pruebas usado por el constructor, i.e., la medida $\mathbb{P}$ representó la probabilidad de construir la prueba p usando el modelo de construcción. Los modelos de construcción representaron restricciones en cómo las pruebas son generadas, por ejemplo, una restricción en las pruebas podría ser evitar repeticiones de claves contiguas en el posicionamiento de las claves. El problema es que varias medidas de probabilidad podrían cumplir estas restricciones, por lo que para elegir la medida de probabilidad que representara al modelo de construcción se escogió la probabilidad que tuviera máxima entropía (ver Definición A.1), que es la probabilidad que captura toda la complejidad de este conjunto de restricciones, o, dicho de otra forma, es la probabilidad que maximiza la incertidumbre que otorga la información de las restricciones.

Para la etapa de resolución de la prueba, se consideró al examinado con un "nivel", que representó el porcentaje de ítems respondidos por conocimiento. Si un examinado que responde una prueba con N ítems, tiene un nivel P_{nivel}, entonces la cantidad de ítems respondidos por conocimiento fue:

$$N_{conoc} = \lfloor N \frac{P_{nivel}}{100} \rfloor$$

Los ítems respondidos por conocimiento se consideraron todos respondidos correctamente. Para elegir los N_{conoc} ítems respondidos por conocimiento, se escogió un subconjunto de indices I_{conoc} de $\{1, 2, ..., N\}$, siguiendo la probabilidad $\mathbb{P}_C$ de máxima entropía en la familia de subconjuntos de $\{1, 2, ..., N\}$ con N_{conoc} elementos, que por la Propiedad A.2 (ver Apéndice A), es la probabilidad uniforme en esta familia de conjuntos. Usando la notación anterior, I_{conoc} es el conjunto de índices de los ítems respondidos por conocimiento.

Terminada la resolución por conocimiento de una prueba $p \in \mathcal{A}^N$ (en los ítems con índices en I_{conoc}), sigue la resolución por estrategia. Una estrategia de resolución (en adelante sólo estrategia) fue definida como una familia F de funciones (ver Definición A.6), donde cada elemento fue una función que toma las respuestas dadas por conocimiento, y entrega respuestas para los ítems faltantes. Se le pidió a esta familia F que estuviera indexada en $\mathcal{P}(\{1, ..., N\}) \setminus \{\{1, ..., N\}\}$, la familia de conjuntos de índices en que puede estar I_{conoc}. Se excluyó $\{1, ..., N\}$ porque si todos los ítems fueran respondidos por conocimiento, no hay uso de estrategia. Se llamará $I_{est} := \{1, ..., N\} \setminus I_{conoc}$, al conjunto de indíces de los ítems respondidos por estrategia. Un elemento de la familia F fue entonces una función $F_{I_{conoc}} : \mathcal{A}^{I_{conoc}} \times \Omega \to \mathcal{A}^{I_{est}}$.

$F_{I_{conoc}}$ posee una componente estocástica Ω porque una estrategia podría no tener preferencia por una opción en ciertos casos. Por ejemplo: Suponga que se quiere modelar la estrategia "evite cualquier tipo de repetición cuando rellene", suponga también que la prueba es de 3 opciones, que las claves de la prueba de 4 ítems son (a, b, a, c) y que el examinado responde por estrategia solo el último ítem (cuya clave es "c"), entonces, para seguir la estrategia debería evitar responder "a" para no causar una repetición. Esto implica que el examinado podría responder indiferentemente "b" o "c" en el último ítem. Para la elección en estos casos se usó Ω, la componente estocástica, para así poder modelar una elección al

azar entre estas dos opciones.

En resumen, el esquema general de modelamiento está dado por cinco parámetros: Número de opciones, número de ítems, modelo de construcción, nivel del examinado y estrategia de resolución. Primero se genera una prueba (con número de opciones y número de ítems dados por los parámetros) desde la probabilidad dada por el modelo de construcción. Segundo, un examinado resuelve la prueba por conocimiento, respondiendo un número de ítems por conocimiento dado por el parámetro de nivel, y los índices de los ítems son generados desde la probabilidad $\mathbb{P}_C$. Tercero, el examinado resuelve el resto de los ítems usando la estrategia, es decir, la función correspondiente al conjunto de indices por conocimiento, toma las respuestas por conocimiento, y genera respuestas al resto de ítems. En total, el proceso (que entrega una prueba y una respuesta a esa prueba) corresponde a una simulación. Cabe notar que, por la aleatoriedad de la estrategia, dos respuestas a una misma prueba, con examinados respondiendo los mismos ítems por estrategia, podrían ser distintas.

Tomando la prueba y la respuesta dada por una simulación, la ganancia de puntos extra (en adelante sólo ganancia) fue calculada como la cantidad de ítems respondidos correctamente por estrategia (en que coincide la clave del ítem y la respuesta del ítem dada por el examinado). También se calculó la ganancia de puntos extra porcentual (en adelante sólo ganancia porcentual) como el porcentaje de puntos extra con respecto al número de ítems respondidos por estrategia. Para simular la distribución de la ganancia y la distribución de la ganancia porcentual para un grupo de parámetros (número de opciones, número de ítems, modelo de construccion, nivel y estrategia), fueron realizadas las simulaciones necesarias para alcanzar una precisión del orden de una décima (para el cálculo y el número de simulaciones usadas en cada estudio ver Apéndice A), calculando la ganancia y la ganancia porcentual en cada simulación. Los modelos de construcción fueron revisados calculando indicadores (a partir del mismo número de simulaciones usado en cada estudio) que caracterizaran a las restricciones subyacentes del estudio correspondiente. Para todos los modelos, se calculó el número de claves por opción de cada prueba; además, para los modelos que involucraron la evitación de repeticiones (simples, dobles o triples), el número de repeticiones (simples, dobles o triples) en cada prueba y el porcentaje de pruebas que tenían alguna repetición (simples, dobles o triples) fue calculado. Al momento de graficar los resultados del número de claves por opción, se presentaron los datos correspondientes a 150 pruebas elegidas aleatoriamente, para obtener una visualización clara de la distribución del número de claves por opción.

Finalmente, se estudiaron conjuntos de modelos de construcción y conjuntos de estrategias, calculando la ganancia y ganancia porcentual en cada par de ellos. Tres análisis fueron hechos para cada conjunto de modelos de construcción y estrategias: 1) Un análisis general donde se estudiaron la ganancia y la ganancia procentual para un número fijo de 4 opciones, un número fijo de 100 ítems (número que permite un cálculo rápido de todos los porcentajes calculados), y 5 niveles de desempeño representativos, 2) un análisis del comportamiento de la ganancia porcentual con respecto al nivel, que consideró niveles donde hay uso de estrategia (0,...,99) y 3) un análisis del comportamiento de la ganancia porcentual con respecto al número de ítems, considerando 5 niveles representativos y número de ítems desde 4 a 100 (considerando sólo los múltiplos de 4 porque uno de los modelos era balancear el número de claves de cada opción).

2.1.2. Modelos de construcción

Los modelos de construcción considerados se basaron en protocolos reportados usados por constructores para posicionar las claves. El primer modelo fue basado en la recomendación randomizar, encontrada en numerosas guías de construcción de pruebas [18, 11, 40]. El segundo modelo fue basado en dos recomendaciones que aparecen recurrentemente en guías de construcción de pruebas: Balancear globalmente [19, 38, 36, 9] y balancear localmente (i.e., evitar repeticiones de 2 claves seguidas [33, 37, 31, 5], de 3 claves seguidas [33, 37, 31, 12, 27], o de 4 claves seguidas [27]). El tercer modelo fue basado en lo que estaría occuriendo cuando los constructores no prestan atención a la distibución de las claves, dejando así una distribución de posiciones sesgada hacia B y C [1]. A continuación se detallan los tres modelos:

- **Modelo Random**: Este modelo fue una probabilidad uniforme sobre el conjunto de secuencias $\{a, b, c, d\}^N$, i.e., entregó igual probabilidad a cada prueba en $\{a, b, c, d\}^N$. Esta medida coincide con la probabilidad de máxima entropía de $\{a, b, c, d\}^N$ sin restricciones (Corolario A.3).
- **Modelo Balanceado Corregido**: Para construir este modelo, dos recomendaciones recurrentes fueron usadas: Balancear globalmente (desde ahora sólo balancear) la posición de la opción correcta (usar la misma cantidad de claves por cada opción en la prueba) y balancear localmente (desde ahora evitar repeticiones), i.e., evitar repeticiones de posición de clave a lo largo de la prueba (evitar que ítems sucesivos tengan la misma clave). Se llama al modelo Corregido porque para construir una prueba sin repeticiones lo que generalmente se hace es corregir una prueba generada al azar. Tres tipos de repeticiones fueron encontradas asociadas a la recomendación de evitación de repeticiones: simples, i.e., 2 claves seguidas en la misma posición [33, 37, 31, 5], dobles, i.e., 3 claves seguidas en la misma posición [33, 37, 31, 12, 27], y triples, i.e., 4 claves seguidas en la misma posición [27]. De las guías educativas que recomiendan evitar repeticiones, la directriz más frecuentemente entregada es evitar repeticiones dobles, por lo tanto, el modelo que se usó aquí fue una probabilidad en el conjunto de las pruebas balanceadas que no tienen repeticiones dobles (de 3 o más alternativas seguidas). La probabilidad de máxima entropía en este caso es la probabilidad uniforme en el conjunto anterior (Corolario A.3).
- **Modelo Sesgo Céntrico**: Este modelo fue construido en base al sesgo céntrico observado en el posicionamiento de claves de constructores que no se fijan en la posición de las opciones de sus ítems, i.e., la tendencia a posicionar más claves en las opciones centrales que en las opciones de los extremos. Este sesgo se observa en el porcentaje promedio del número de claves en cada opción a nivel de la prueba, por lo que, si se denota (P_a, P_b, P_c, P_d) a los porcentajes promedio de cada opción, la probabilidad $\mathbb{P}_{sesgo\text{-}centrico}$ del modelo debería cumplir que:

  $$\forall alternativa \in \{a, b, c, d\} : \frac{N_{alternativa}}{N} \cdot 100 = P_{alternativa}$$

  Luego, escogiendo la probalidad de máxima entropía que cumple esta restricción, se tiene que, por el Corolario A.5, $\mathbb{P}_{sesgo\text{-}centrico}$ es el producto de N probabilidades discretas iguales sobre el conjunto $\{a, b, c, d\}$.

Para calcular los porcentajes $P_{alternativa}$, fueron usados datos reales de distribuciones de claves (conjuntos de ítems de 4 opciones) reportadas como sesgadas hacia el centro por sus autores. Diez casos proviniendo de cinco estudios fueron recuperados [1, 28, 29, 22, 30], usando una base de datos asociada a una revisión sistemática previa [25]. Suponiendo que los porcentajes encontrados en los datos vienen de una única distribución (la del modelo de sesgo que describimos anteriormente), los porcentajes de esta distribución fueron estimados promediando los porcentajes de claves en cada opción encontrados en las distribución de claves de los datos. Los porcentajes encontrados fueron:

$$P_a = 19{,}4\,\%, P_b = 29{,}5\,\%, P_c = 28{,}8\,\%, P_d = 22{,}2\,\%$$

Luego, la probabilidad discreta fue dada por:

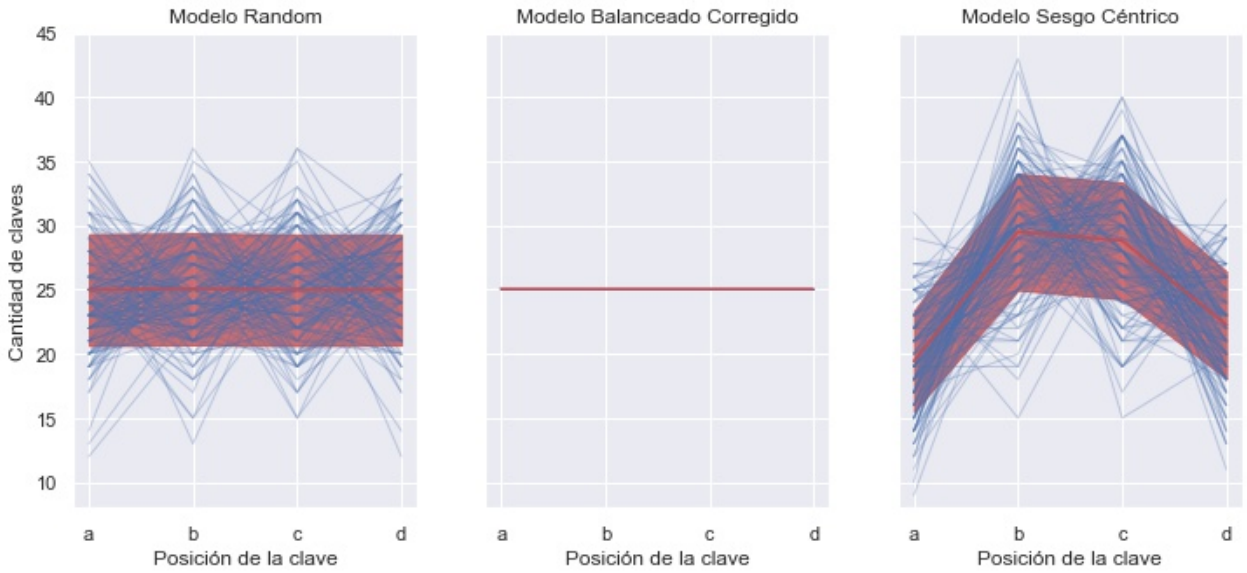
$$\mathbb{P}_{centrica}(\{alternativa\}) = (P_{alternativa}/100), \forall alternativa \in \{a, b, c, d\}$$

Con una probabilidad mayor en las alternativas centrales (b y c) que las extremales (a y d). Si $p = (p_1, p_2, ..., p_N)$ es una prueba cualquiera, su probabilidad está dada por:

$$\mathbb{P}_{sesgo\text{-}centrico}(p) = \prod_{i=1}^{N} \mathbb{P}_{centrica}(p_i)$$

20.000 simulaciones fueron realizadas para verificar estos modelos (ver Apéndice B). Con respecto al número de claves en cada opción, acorde a lo esperado, se observa en la Figura 2.1 que: 1) el modelo Balanceado Corregido obtuvo distribuciones de claves que fueron perfectamente balanceadas entre opciones, 2) el modelo Random, obtuvo el mismo número promedio de claves en cada opcion, pero que este número presentó una alta variabilidad, tal que obtener una distribución perfectamente balanceada fue muy poco probable, y 3) el modelo Sesgo Céntrico tuvo mayor número promedio de claves en opciones céntricas, con una variabilidad similar a la del modelo Random en cada alternativa. También se pudo ver en la Figura 2.2 que ninguna prueba del modelo Balanceado Corregido tuvo repeticiones dobles, contrario a los modelos Random y Sesgo Céntrico que casi siempre tuvieron alguna repetición. Finalmente, se puede ver en la Figura 2.3 que la cantidad de repeticiones del modelo Random y del modelo Sesgo Céntrico fueron similares. En resumen, los modelos de construcción condujeron a la construcción de pruebas cuyo posicionamiento de claves correspondió efectivamente a lo que se esperaba obtener a partir de los supuestos teóricos.

Figura 2.1: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 2.2: Porcentaje de pruebas con repeticiones dobles

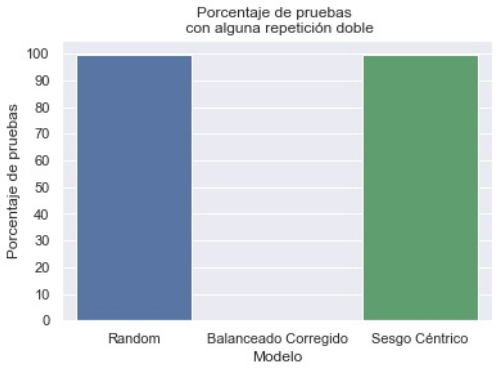
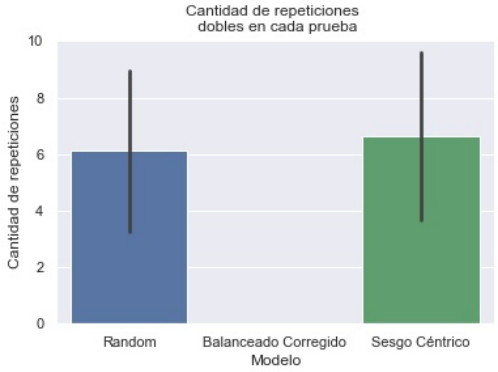


Figura 2.3: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

2.1.3. Estrategias de resolución

A continuación se detallan dos estrategias que fueron encontradas documentadas y una estrategia usada como control:

- **Estrategia Random**: Esta estrategia fue escogida como una estrategia teórica control y corresponde a responder de manera aleatoria y uniforme cada ítem del conjunto de

ítems que se responden por estrategia.

- **Estrategia Pura C**: La estrategia Pura C, que es de bajo nivel cognitivo, fue reportada en una entrevista informal en [8], y aunque no se ha demostrado que la estrategia sea usada en pruebas reales, un fuerte indicio puede ser encontrado en [1], donde encontraron que rachas de respuestas iguales al final de una prueba tienen mayor frecuencia al centro ("*b*" y "*c*") que en los extremos ("*a*" y "*d*"). Esta estrategia corresponde a seleccionar "*c*" como respuesta para todos los ítems que se responden por estrategia.
- **Estrategia Underdog**: La estrategia Underdog, que es de más alto nivel cognitivo, fue descrita en [3] como una conducta estratégica optimizada para aprovechar el balanceo de las pruebas y fue definida como:

  1. Responda todas las preguntas que pueda responder por conocimiento.
  2. Cuente la frecuencia de cada posición entre sus respuestas.
  3. Seleccione la posición con menor frecuencia ("the underdog position"). En caso de empate, elija cualquiera de ellas.
  4. Responda todas las preguntas que falta con la posición seleccionada.

  La estrategia corresponde entonces a responder todos los ítems con una misma opción, la opción que haya tenido menos respuestas en la fase de resolución por conocimiento. Si dos opciones tuvieran el mínimo número de respuestas, entonces se debe escoger aleatoria y uniformemente entre las dos, y responder todos los ítems faltantes con esta opción.

2.1.4. Análisis

Primero, para cada modelo de construcción y para cada estrategia, la ganancia y la ganancia porcentual fue simulada para pruebas con un número fijo de 100 ítems y examinados con 5 diferentes niveles: 60 % (representando el nivel mínimo de aprobación en Chile), 40 % (representando un nivel bajo de 20 % menos que el nivel de aprobación), 80 % (representando un nivel alto de 20 % mas que el nivel de aprobación), 10 % (representando un nivel muy bajo) y 90 % (representando un nivel muy alto). La ganancia fue comparada en cada nivel entre los distintos pares (modelo, estrategia) usando un t-test. La diferencia entre las ganancias, así como el tamaño y el grado de significancia de los efectos encontrados fueron analizados (usando el coeficiente d de Cohen y el p-valor para estos dos últimos respectivamente). Los coeficientes d de Cohen fueron interpretados como: Despreciable ($d < 0.2$), pequeño ($0.2 \leq d < 0.5$), mediano ($0.5 \leq d < 0.8$), grande ($0.8 \leq d < 1.3$) y muy grande ($1.3 \leq d$). La ganancia porcentual fue analizada descriptivamente para comparar las ganancias entre distintos niveles de desempeño.

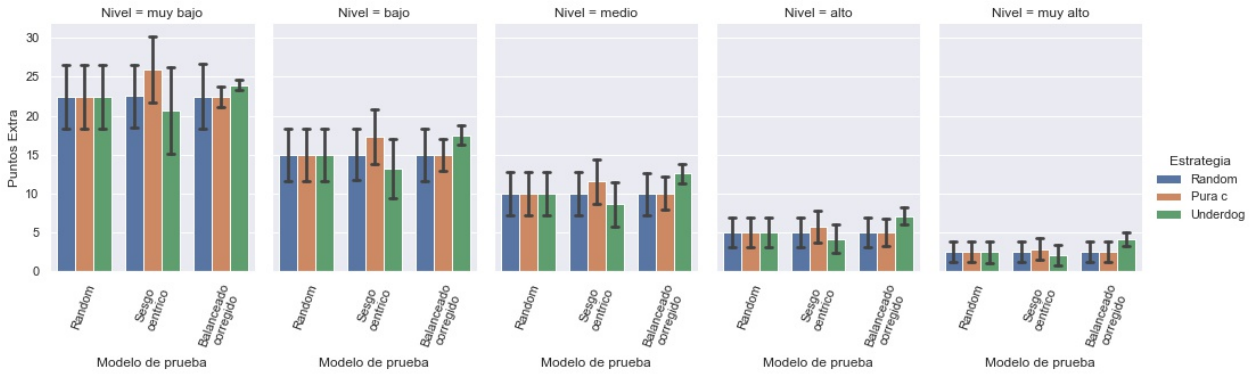
Segundo, la ganancia porcentual fue simulada para un número fijo de 100 ítems en los niveles en que no todas las preguntas fueron respondidas por conocimiento, i.e., $nivel \in \{0, ..., 99\}$. La ganancia porcentual fue analizada descriptivamente con respecto al nivel.

Tercero, la ganancia porcentual fue simulada para los 5 niveles definidos anteriormente, en los números de ítems que admitieran pruebas balanceadas, i.e., $N \in \{4{\cdot}1, 4{\cdot}2, ..., 4{\cdot}25 = 100\}$. La ganancia porcentual fue analizada descriptivamente con respecto al número de ítems.

2.2. Resultados

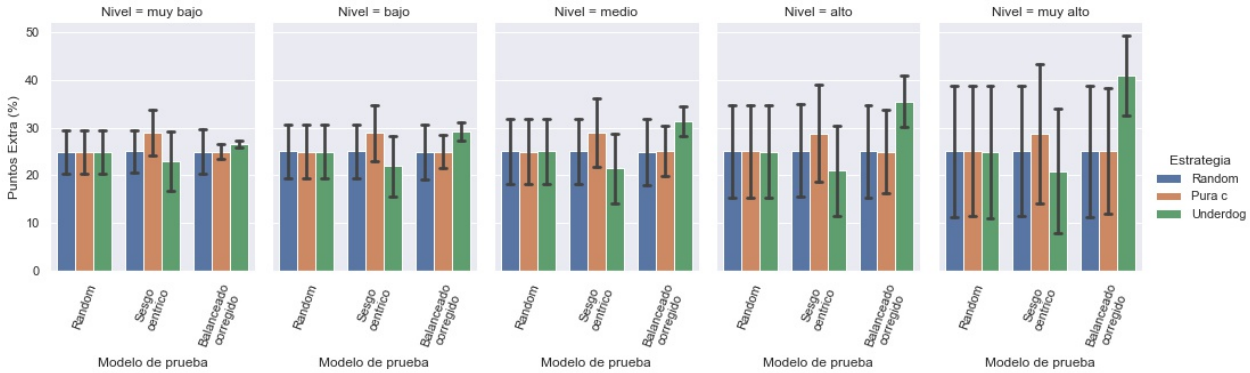
La ganancia y la ganancia porcentual obtenidas en 100 ítems y 5 niveles son presentadas en las Figuras 2.4 y 2.5, respectivamente. En el modelo sesgo céntrico, la ganancia de la estrategia Pura C fue significativamente mayor (con tamaños de efecto desde pequeño a grande) a la ganancia de la estrategia control Random (Δ_{10}=3.5, d_{10}=0.8; Δ_{40}=2.3, d_{40}=0.7; Δ_{60}=1.6, d_{60}=0.6; Δ_{80}=0.7, d_{80}=0.4; Δ_{90}=0.4, d_{90}=0.3; p-values < 0.0001) y, en el modelo Balanceado Corregido, la ganancia de la estrategia Underdog fue significativamente mayor (con tamaños de efecto desde mediano a muy grande) a la ganancia de la estrategia control Random (Δ_{10}=1.5, d_{10}=0.5; Δ_{40}=2.5, d_{40}=1.0; Δ_{60}=2.6, d_{60}=1.2; Δ_{80}=2.1, d_{80}=1.4; Δ_{90}=1.6, d_{90}=1.4; p-values < 0.0001), indicando que existe una ventaja significativa en el uso de estrategias de resolución basada en la posición de las opciones en ciertos protocolos de construcción de pruebas.

Figura 2.4: Ganancias obtenidas para el Estudio 1.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 2.5: Ganancias porcentuales obtenidas para el Estudio 1.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

En el modelo Random, las ganancias de las tres estrategias no fueron distintas (Random v/s Pura C: p-values > 0.1; Random v/s Underdog: p-values > 0.5; Pura C v/s Underdog: p-values > 0.4), confirmando que el modelo Random es un modelo de contrucción control

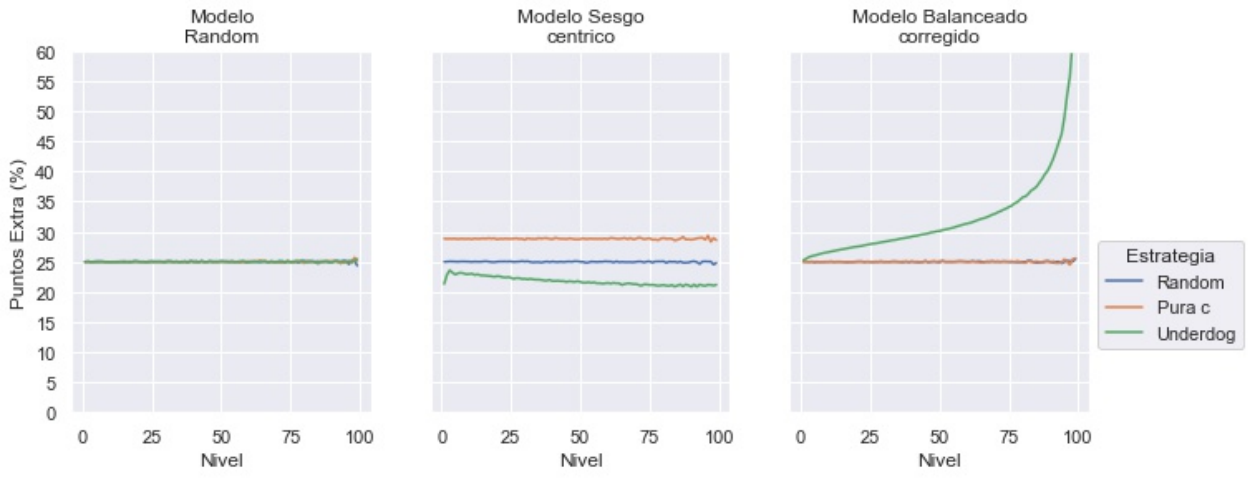
para el estudio de distintas estrategias (ver Corolario A.8) y que en este modelo las estrategias no permiten la obtención de puntajes extras que no dependan del conocimiento. Asimismo, la ganancia de la estrategia Random fue similar en los tres modelos de prueba (Random v/s Sesgo Céntrico: p-values > 0.1; Random v/s Balanceado Corregido: p-values > 0.1; Sesgo Céntrico v/s Balanceado Corregido: p-values > 0.1), confirmando que la estrategia Random es una estrategia control para el estudio de la ganancia en distintos modelos (ver Corolario A.9).

De modo interesante, en el modelo Sesgo Céntrico la ganancia de la estrategia Underdog no sólo fue significativamente menor a la ganancia de la estrategia Pura C (Δ_{10}=5.3, d_{10}=1.1; Δ_{40}=4.1, d_{40}=1.1; Δ_{60}=3.0, d_{60}=1.0; Δ_{80}=1.5, d_{80}=0.8; Δ_{90}=0.8, d_{90}=0.6; p-values < 0.0001), sino que también fue significativamente menor (con tamaños de efecto desde pequeño a mediano) a la estrategia control Random (Δ_{10}=1.8, d_{10}=0.4; Δ_{40}=1.9, d_{40}=0.5; Δ_{60}=1.4, d_{60}=0.5; Δ_{80}=0.8, d_{80}=0.4; Δ_{90}=0.4, d_{90}=0.3; p-values < 0.0001), mostrando que una misma estrategia puede sacar ventaja en un modelo de construcción pero hacer perder puntos en otro modelo. Por otro lado, en el modelo Balanceado Corregido, la ganancia de Pura C no fue distinta de la ganancia de la estrategia control Random (p-values > 0.1).

En la Figura 2.5, se observa que la ganancia porcentual en la estrategia Pura C es similar en todas las categorías de desempeño. En cambio, la ganancia porcentual de la estrategia Underdog aumenta cuando el nivel aumenta, indicando que los estudiantes de más alto desempeño sacan mayor provecho de esta estrategia.

En la Figura 2.6, se presenta la ganancia porcentual para los distintos niveles de desempeño posibles.

Figura 2.6: Ganancia porcentual para niveles en $\{0, ..., 99\}$.



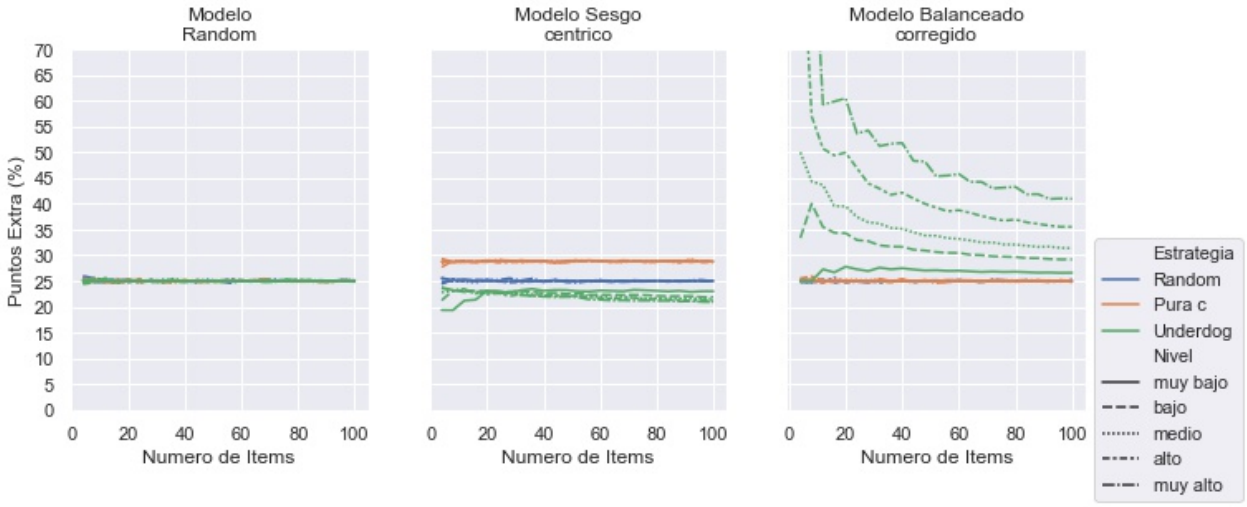
Nota: En el primer gráfico las 3 líneas están superpuestas. En el segundo gráfico las líneas de Pura C y Random están superpuestas.

Confirmando la observación de la ganancia porcentual en 5 niveles, se puede ver en la Figura 2.6 que la ganancia porcentual en Pura C es constante con respecto al nivel y que la ganancia porcentual de Underdog es creciente con respecto al nivel (de hecho, con un creci-

miento aproximadamente exponencial). Además, se observa que la desventaja de Underdog en Sesgo Céntrico es creciente con respecto al nivel, pero con una pendiente pequeña.

La ganancia porcentual para distinto número de ítems se presenta en la Figura 2.7. Se observa que el número de ítems no tiene influencia en la ganancia porcentual de la estrategia Pura C, pero que la ganancia porcentual de la estrategia Underdog aumenta drásticamente en pruebas cuyo número de ítems es bajo. Cabe destacar que esta ganancia porcentual es siempre mayor a la ganancia de la estrategia Pura C para todos los niveles excepto el muy bajo.

Figura 2.7: Ganancia porcentual para número de ítems en $\{4, 8, ..., 100\}$



2.3. Discusión

A partir de los resultados del Estudio 1, se puede observar que el uso de estrategias de resolución basadas en la posición de las opciones permite la obtención de puntos extras en pruebas de selección múltiple. Esta ventaja en el desempeño, que se obtiene sin la necesidad de contar con un conocimiento disciplinar, existe tanto en pruebas con número pequeño de ítems como en pruebas con un gran número de ítems, y es posible de obtener por los estudiantes de todo nivel.

Los puntos extras que se puedan obtener inyectan ruido no deseado en la medición; representan una amenaza para la validez de los resultados de la prueba 1) en los casos en que la posición de las opciones correctas no haya sido cuidada siguiendo algún protocolo recomendado en guías de construcción (modelo Sesgo Céntrico), y 2) incluso si se cuida la posición de las opciones correctas, cuando se haya hecho siguiendo la recomendación de balancear y evitar repeticiones.

En los resultados del Estudio 1, destaca la ventaja porcentual de la Estrategia Pura C (con tamaños de efecto desde pequeños hasta incluso grandes), que muestra que el modelo Sesgo Céntrico conlleva potenciales problemas de validez de los resultados. Siendo esta

ventaja estable para examinados con distintos niveles de desempeño, no pareciera conllevar este modelo problemas de equidad. La ventaja es también similar para pruebas con distintos números de ítems. Debido a los problemas de validez señalados, se concluye que no es recomendable despreocuparse de la posición de las opciones correctas en la fase de construcción de pruebas de selección múltiple.

Por otro lado, destaca la ventaja porcentual de la estrategia Underdog en el modelo Balanceado Corregido. Asociada a un tamaño de efecto variable (desde medio hasta incluso muy grande), esta ventaja indica que el modelo Balanceado Corregido también conlleva problemas de validez. Conlleva además problemas de equidad, pues la ventaja porcentual varía según el nivel de desempeño, siendo mayor mientras mayor es el nivel de desempeño del examinado. Finalmente, la ventaja porcentual también fue mayor mientras menor fue el número de ítems, por lo que se puede concluir que no es recomendable seguir la recomendación de balancear y evitar repeticiones, y que evitarlo es particularmente importante cuando se construyen pruebas de pocos ítems.

En términos cuantitativos, se entiende por qué la estrategia Pura C saca ventaja del modelo Sesgo Céntrico, y es porque la ganancia porcentual promedio de la estrategia Pura C está dada por la frecuencia de la opción C en el sesgo céntrico. Esto mismo explicaría la desventaja de Underdog observada en el Sesgo Céntrico, la mayoría de respuestas observadas en general son las de mayor frecuencia en el sesgo céntrico, por lo que, como Underdog elige la respuesta de menor frecuencia, el puntaje porcentual estará dado por la menor frecuencia en el sesgo céntrico, que es menor que 25 %, es decir, menor que Random.

Lo que no es tan claro es de donde viene la alta ganancia de la estrategia Underdog en el modelo Balanceado Corregido, es decir, si esta estrategia saca provecho de ambas recomendaciones (balenceo global, balanceo local) o de alguna en particular (esta pregunta será aclarada en el Estudio 2). Aunque el fenómeno de variación en la ganancia porcentual con respecto al nivel de desempeño (que también había sido observado en [3] para pruebas de alta consecuencia) y al número de ítems observado para Underdog en Balanceado Corregido sí puede entenderse: se puede entender como una ganancia porcentual mayor mientras más información se tenga de la prueba (o bien se tiene menos items y el mismo nivel, o bien se tiene mayor porcentaje de respuestas por conocimiento, i.e., mayor nivel).

Los resultados del Estudio 1 también permiten verificar que el modelo Random no deja lugar a ninguna ventaja por estrategia, pues las ganancias de todas las estrategias en el modelo Random son iguales (Corolario A.8). Esto implica entonces que, a priori, este modelo sería el único que no pondría en riesgo la validez de los resultados. Por otra parte, también se pudo verificar la calidad de la estrategia Random como control, es decir, que para cualquier modelo su ganancia fuera la misma (Corolario A.9). Con ello, se justifica que sea posible afirmar que una estrategia tiene una ventaja (o desventaja) en un modelo si su ganancia es mayor (o menor) a la ganancia de la estrategia Random en ese modelo.

El hecho de que la recomendación de balanceo no pareciera ser adecuada es un resultado potencialmente preocupante, considerando que numerosas guias educativas proporcionan directrices de balanceo. Este hallazo merece ser estudiado con mayor detención. Es importante señalar que el balanceo que los constructores aplican no siempre es exacto, y que algunas recomendaciones de balanceo contemplan ubicar un número similar, no idéntico, de claves

en cada posición de opción (el balanceo se llama aproximativo en este caso). También cabe mencionar que el modelo Balanceado Corregido no estaba basado en un simple balanceo a nivel de prueba, sino que además contemplaba un balanceo local. Como no es claro que la recomendación de balancear aproximativamente deje lugar a que alguna estrategia pueda sacar beneficio, y como vale la pena preguntarse cual es el efecto exacto del balanceo global y local, en el siguiente estudio se analizan modelos basados en las recomendaciones de balanceo aproximativo y balanceo global y local por separado, además de si es posible sacar más ventaja del modelo Balanceado Corregido con una estrategia más compleja que Underdog.

Capítulo 3

Estudio 2: Ganancia del uso de la estrategia Underdog frente al balanceo local, balanceo exacto y balanceo aproximativo

Este estudio tuvo tres objetivos: 1) Evaluar cuanta ventaja saca la estrategia Underdog del balanceo exacto y de la evitación de repeticiones, aplicado por separado (en comparación a aplicado en conjunto como en el modelo de construcción de Balanceado Corregido), 2) evaluar si es posible encontrar una estrategia más efectiva que Underdog para sacar ventaja de los modelos de Balanceado, y 3) evaluar si la estrategia Underdog sigue sacando ventaja cuando el balanceo es aproximativo y no exacto.

3.1. Método

El diseño general de este estudio fue el mismo que en el Estudio 1.

3.1.1. Modelos de construcción

- **Modelo Balanceado Corregido**: Definido en el Estudio 1. Este modelo se desagrupa en los modelos siguientes, que están basados en las dos directrices subyacentes al Balanceado Corregido: Balancear de forma exacta y evitar repeticiones dobles.
- **Modelo Random Corregido Por Repeticiones Dobles**: Para construir este modelo, la recomendación de evitar repeticiones dobles fue usada. El modelo será una probabilidad en el conjunto de las pruebas que no tienen repeticiones dobles. La probabilidad de máxima entropía en este caso será la probabilidad uniforme en este conjunto (Corolario A.3).

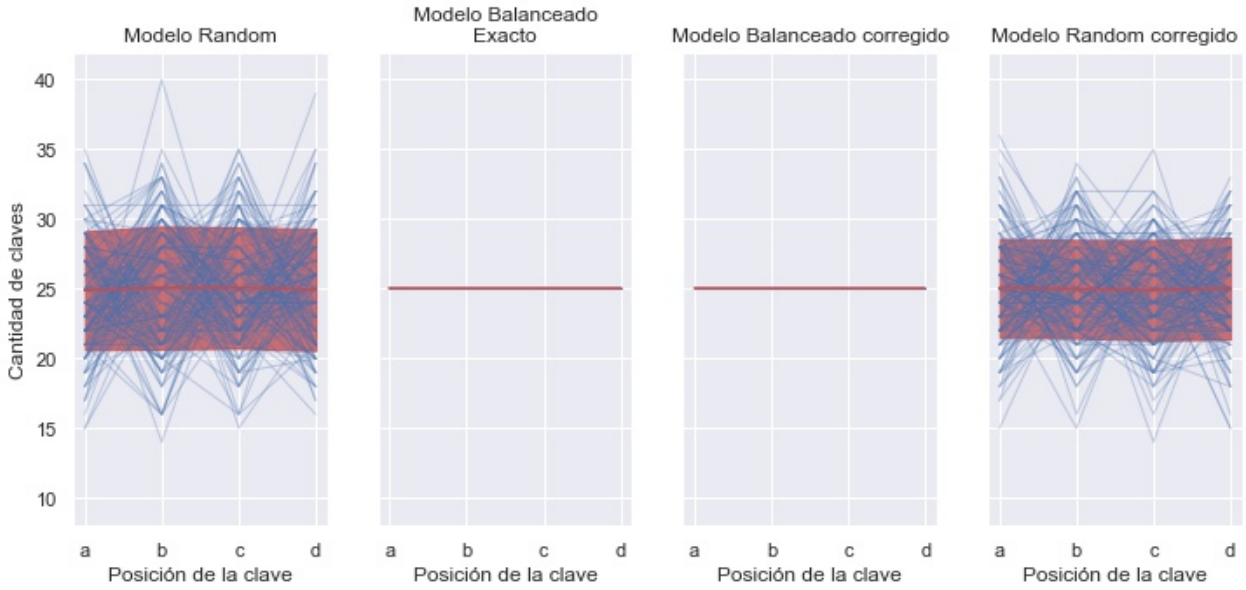
- **Modelo Balanceado Exacto**: Para describir este modelo, la recomendación de balancear fue usada. Esta recomendación, en el caso que se respondieran todas las preguntas al azar, daría como resultado una probabilidad de 1/4 para cada opción en cada prueba, lo que se podría interpretar como un tipo de justicia, además, evitaría que alguna de las alternativas no apareciera (o apareciera muy frecuentemente) como clave, lo que pudiera confundir al examinado. El modelo es una probabilidad en el conjunto de las pruebas balanceadas. La probabilidad de máxima entropía en este caso es la probabilidad uniforme en este conjunto (Corolario A.3).

- **Modelo Balanceado Aproximativo**: Este modelo fue basado en la recomendación de balancear aproximadamente, i.e., cuidar que el número de claves en cada opción no sea muy distinto a una distribución exactamente balanceada. Para cuantificar la cercanía a una distribución balanceada de forma exacta, usamos un criterio basado en un parámetro que llamaremos delta. Este criterio corresponde a calcular la diferencia máxima en el número de claves que hay entre opciones, y luego, tomar esta diferencia como un porcentaje con respecto al número de opciones. El modelo Balanceado Aproximativo δ serán entonces todas las pruebas cuyo criterio sea menor o igual a δ. En este sentido, una prueba balanceada de forma exacta es una prueba con criterio delta 0.

  El criterio delta fue calculado en datos previos de pruebas de 4 opciones que fueron declaradas aproximadamente balanceadas por sus autores. 24 pruebas proviniendo de 6 estudios fueron recuperadas [19, 39, 38, 1, 13, 26]. Los delta calculados fueron en promedio 4.4 ($\pm$1,7), con un mínimo de 1.3 y un máximo de 7.9. Consideramos entonces delta en el rango $\{1, ..., 8\}$. Un modelo Balanceado Aproximativo (para δ) es una probabilidad en el conjunto de las pruebas cuyo criterio delta sea menor o igual a δ. La probabilidad de máxima entropía en este caso será la probabilidad uniforme en este conjunto (Corolario A.3).

Simulaciones fueron realizadas para verificar los modelos Random Corregido por repeticiones dobles, Balanceado Exacto y Balanceado Corregido, y para compararlos con el modelo Random del Estudio 1. Con respecto al número de claves en cada opción, acorde a lo esperado, se observa en la Figura 3.1 que en los modelos Balanceado y Balanceado Corregido no hay variabilidad en el número de claves por opción. Los modelos Random y Random Corregido por repeticiones dobles tienen una variabilidad similar en el número de claves por opción. En la Figura 3.2, se puede ver que los modelos corregidos no tienen repeticiones dobles, lo que es contrario a los modelos no corregidos, que si tienen repeticiones dobles, y en la 3.3 se puede ver que la cantidad de repeticiones dobles en los modelos Random y Random Corregido por repeticiones dobles es similar.

Figura 3.1: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 3.2: Porcentaje de pruebas con repeticiones dobles

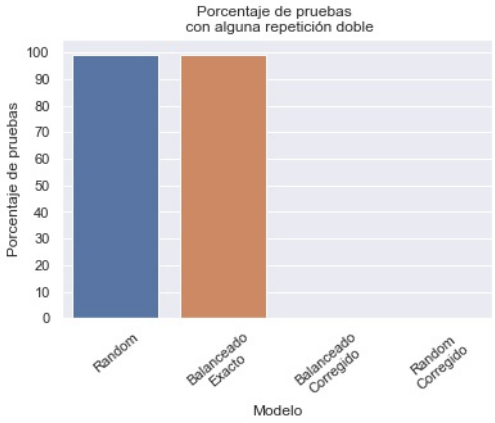
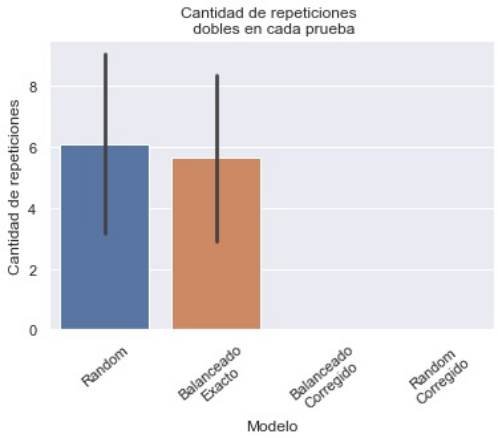
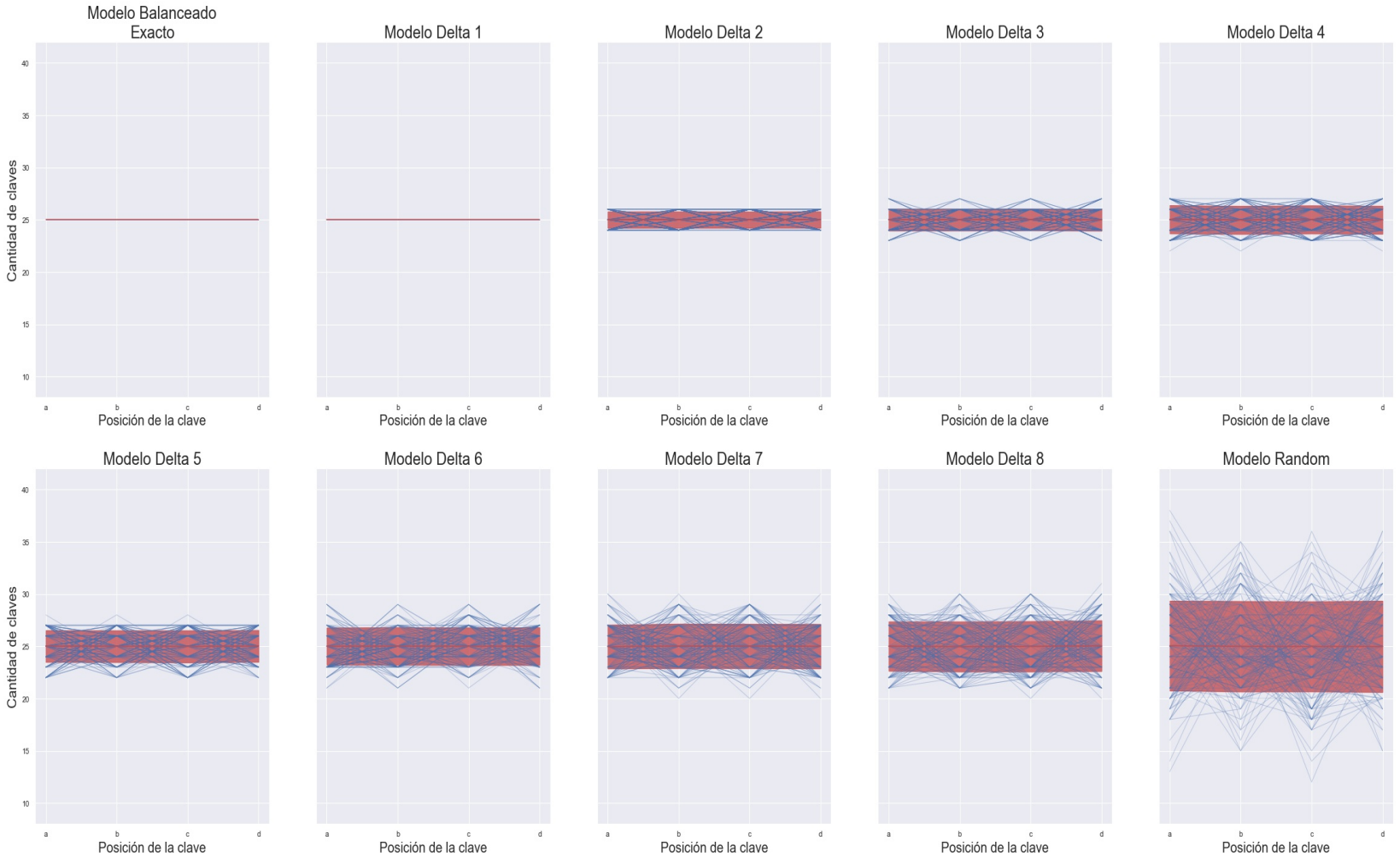


Figura 3.3: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 3.4: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 3.5: Porcentaje de pruebas con repeticiones dobles

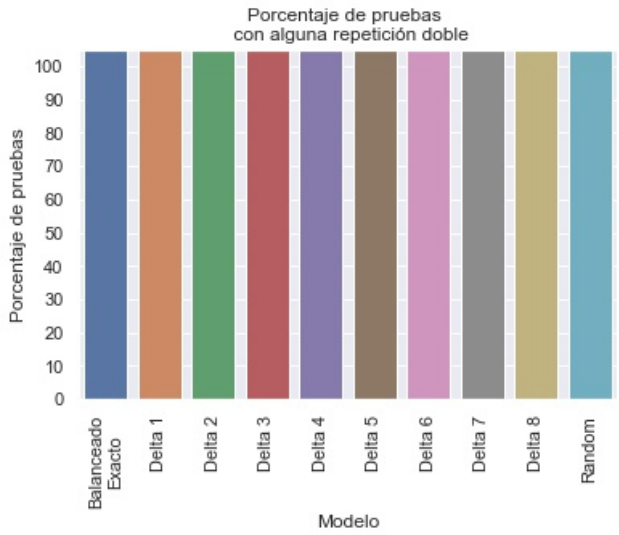
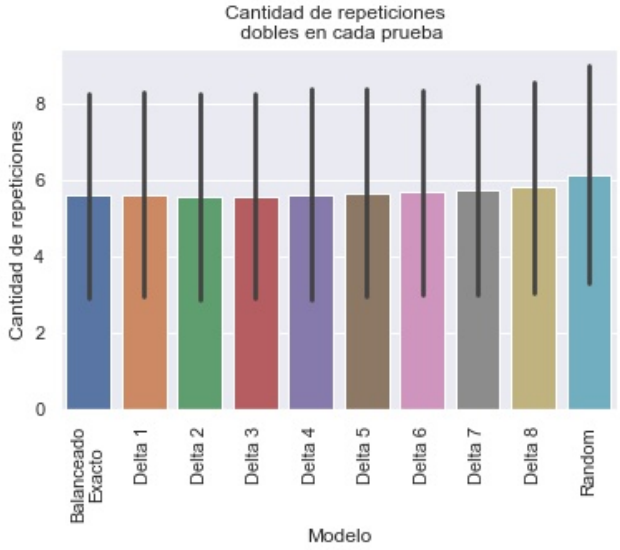


Figura 3.6: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Los modelos Balanceado Aproximativo también fueron verificados con 10.000 simulaciones, y comparados con el modelo Random del Estudio 1 y Balanceado Exacto. En la Figura 3.4, se puede observar que los modelos Aproximativo Delta tienen varianza creciente en delta y que sus varianzas son todas mayores que el modelo Balanceado Exacto y menores que el modelo Random, lo que coincide con la idea de balanceo aproximativo y apoya la decisión del parámetro delta para esta recomendación. Por otra parte, en la Figura 3.5 se ve que todos los modelos tienen casi siempre repeticiones dobles y en la Figura 3.6 se observa que los porcentajes de estas repeticiones dobles son similares en todos los modelos, mostrando que los modelos delta afectan al número de claves por opción y no a las repeticiones. En resumen, los indicadores coinciden con lo esperado a partir de los supuestos teóricos.

3.1.2. Estrategias de resolución

- **Estrategia Underdog**: Definida en el Estudio 1.
- **Estrategia evitar repeticiones dobles**: Esta estrategia corresponde a elegir respuestas que saquen ventaja de pruebas que no tienen repeticiones dobles. Para ello, los ítems son respondidos evitando crear una repetición doble con las respuestas actuales (las respuestas por conocimiento y las respuestas por estrategia hasta el momento). En caso de que haya más de una opción que no cree una repetición doble, se elige aleatoria y uniformemente entre los candidatos.
- **Estrategia Underdog Mejorado**: Esta estrategia corresponde a elegir respuestas que saquen ventaja de pruebas que están balanceadas de forma exacta y que además no tienen repeticiones dobles. Está basada en la estrategia Underdog, que es una estrategia de ganancia maximal para el modelo de Balanceo Exacto (ver Corolario A.12). La idea es que, cuando no es posible seleccionar la opción que se eligiera con la estrategia Underdog (la opción con menor frecuencia de uso) porque seleccionarla crearía una repetición doble, se selecciona la opción que tenga menor frecuencia dentro del conjunto de respuestas que no causan repeticiones dobles, y en el caso de que hayan 2 o más candidatos se elige aleatoria y uniformemente entre ellos.

3.1.3. Análisis

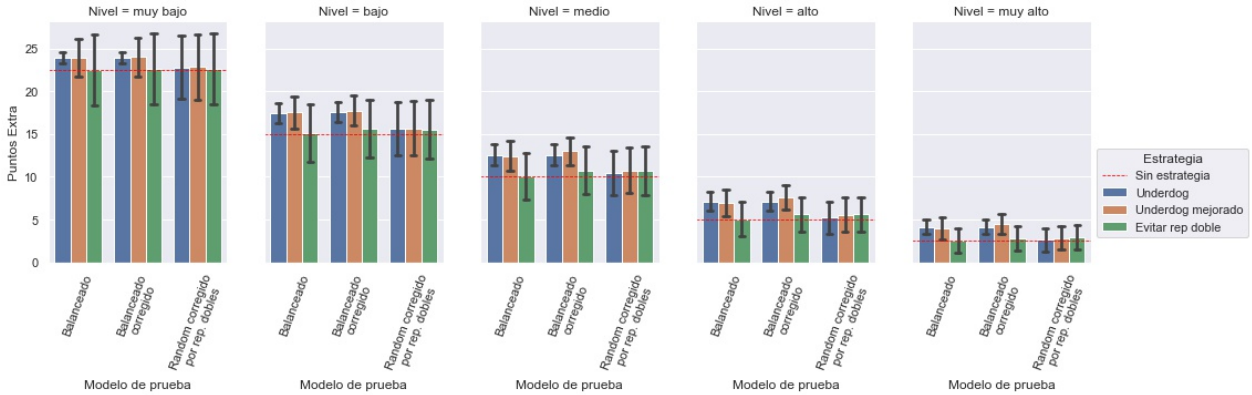
Se hicieron los mismos análisis que en el Estudio 1 para: 1) los modelos de construcción Balanceado, Random Corregido por repeticiones dobles, y Balanceado Corregido, y las estrategias Underdog, evitar repeticiones dobles y Underdog Mejorado; y 2) los modelos de construcción Balanceado Exacto, Balanceado aproximativo (con $\delta \in \{1, ..., 8\}$) y la estrategia Underdog.

3.2. Resultados

Para los modelos Balanceado, Balanceado Corregido y Random Corregido por repeticiones dobles, y las estrategias Random, Underdog, Underdog Mejorado y Evitar Repeticiones Dobles, la ganancia y la ganancia porcentual obtenidas en 100 ítems y 5 niveles son presentadas en las Figuras 3.7 y 3.8 respectivamente.

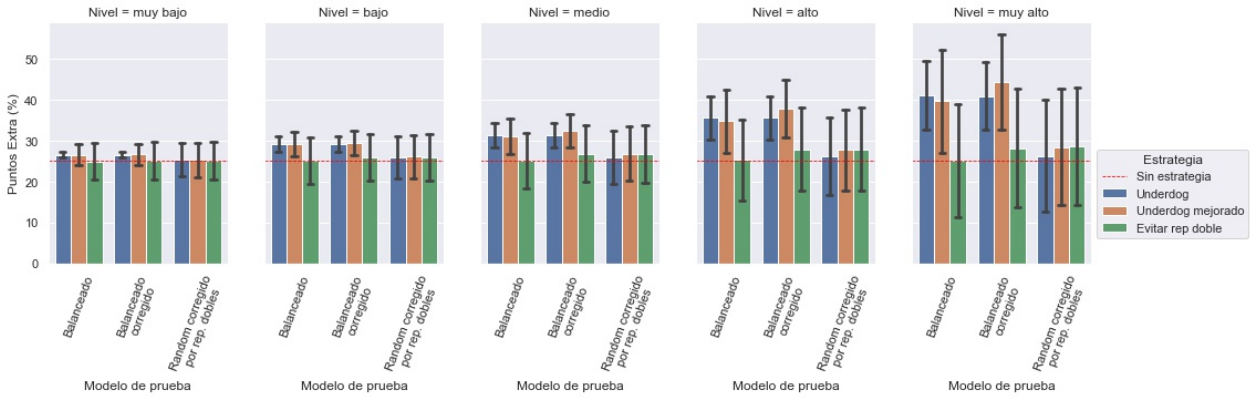
En el modelo Balanceado Exacto, la ganancia obtenida por la estrategia Underdog fue significativamente mayor que la ganancia obtenida por la estrategia control Random (Δ_{10}=1.5, d_{10}=0.5; Δ_{40}=2.5, d_{40}=1.0; Δ_{60}=2.5, d_{60}=1.2; Δ_{80}=2.1, d_{80}=1.4; Δ_{90}=1.6, d_{90}=1.4; p-values < 0.0001). Esta ganancia fue idéntica para los modelos Balanceado Exacto y Balanceado Corregido (p-values > 0.2). En el modelo Random Corregido Por Repeticiones Dobles, la ganancia de Underdog fue similar (tamaños de efecto despreciables) a la ganancia de la estrategia Random (Δ_{10}=0.3; Δ_{40}=0.5; Δ_{60}=0.4; Δ_{80}=0.3; Δ_{90}=0.1;p-values<0.0001; d_{10}, d_{40}, d_{60}, d_{80}, d_{90}=0.1). Estos resultados indican que la ventaja de Underdog reportada en el Es-

Figura 3.7: Ganancias obtenidas para el Estudio 2.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 3.8: Ganancias porcentuales obtenidas para el Estudio 2.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

tudio 1 proviene totalmente del hecho de que esta estrategia de resolución saca provecho de las distribuciones balanceadas de claves en las pruebas, y no de la evitación de repeticiones.

Comparando Underdog con Underdog Mejorado, la ganancia de Underdog Mejorado fue levemente mayor (con tamaños de efecto pequeños) o igual que la ganancia de Underdog en el modelo Balanceado Corregido (Δ_{40}=0.2, d_{40}=0.1; Δ_{60}=0.4, d_{60}=0.3; Δ_{80}=0.4, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.3; p_{40}, p_{60}, p_{80}, $p_{90} < 0.0001$, $p_{10} > 0.1$). pero en el modelo Balanceado Exacto, la ganancia de Underdog Mejorado fue levemente menor (con tamaños de efecto despreciables) o igual que la ganancia de Underdog (Δ_{60}=0.1, d_{60}=0.1; Δ_{80}=0.1, d_{80}=0.1; Δ_{90}=0.1, d_{90}=0.1; p_{60}, p_{80}, $p_{90} < 0.0001$, p_{10}, $p_{40} > 0.1$). Estos resultados indican que es posible mejorar la ventaja que tiene Underdog en Balanceo Corregido, aunque la ventaja encontrada es pequeña. Además, se confirma que la estrategia Underdog es máximal con respecto al Balanceo Exacto (ver Corolario A.12).

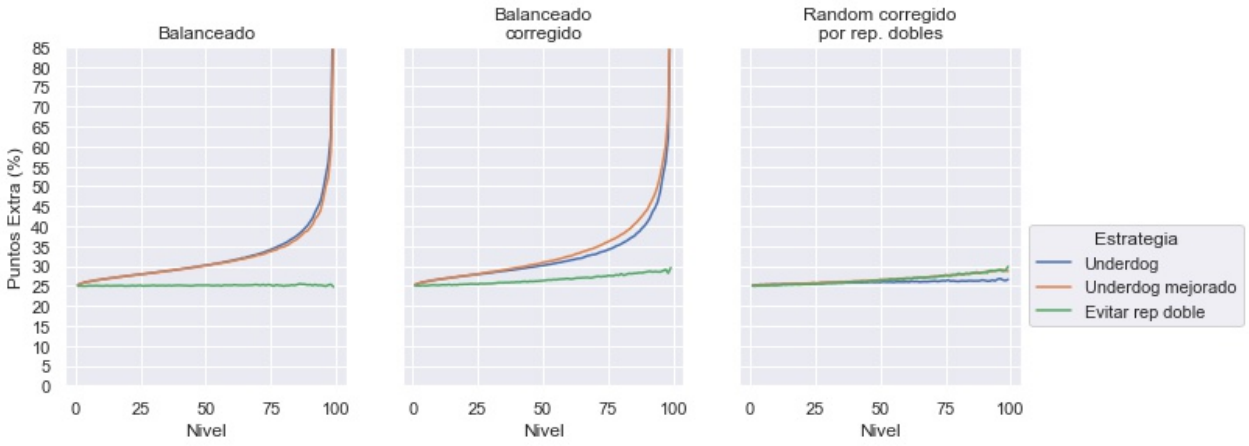
En los modelos Balanceado Corregido y Random Corregido por repeticiones dobles la ganancia de la estrategia Evitar Repeticiones Dobles es significativa pero solo es levemente mayor (tamaños de efecto desde despreciables hasta pequeños en ambos casos) o igual que

la ganancia de la estrategia Random (Balanceado Corregido: Δ_{10}=0.1, d_{10}=0.0; Δ_{40}=0.6, d_{40}=0.2; Δ_{60}=0.6, d_{60}=0.2; Δ_{80}=0.5, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.2; p-values < 0.0001 ; Random Corregido por repeticiones dobles: Δ_{40}=0.6, d_{40}=0.2; Δ_{60}=0.7, d_{60}=0.3; Δ_{80}=0.6, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.2; p_{40},p_{60},p_{80},p_{90} < 0.0001, p_{10}=0.1). indicando que la estrategia basada en evitar repeticiones dobles puede sacar una ventaja significativa aunque esta sea mas bien pequeña.

Se observa en la Figura 3.8 que en los modelos Balanceado Exacto y Balanceado Corregido la ganancia porcentual de las estrategias Underdog y Underdog Mejorado son crecientes con respecto al nivel del examinado, indicando que los examinados con mayor nivel de desempeño pueden sacar mayor ventaja de las estrategias basadas en el balanceo. Por otra parte, en los modelos Balanceado Corregido y Random Corregido por repeticiones dobles, la ganancia porcentual de las estrategias Underdog y evitar repeticiones dobles es similar en todos los niveles.

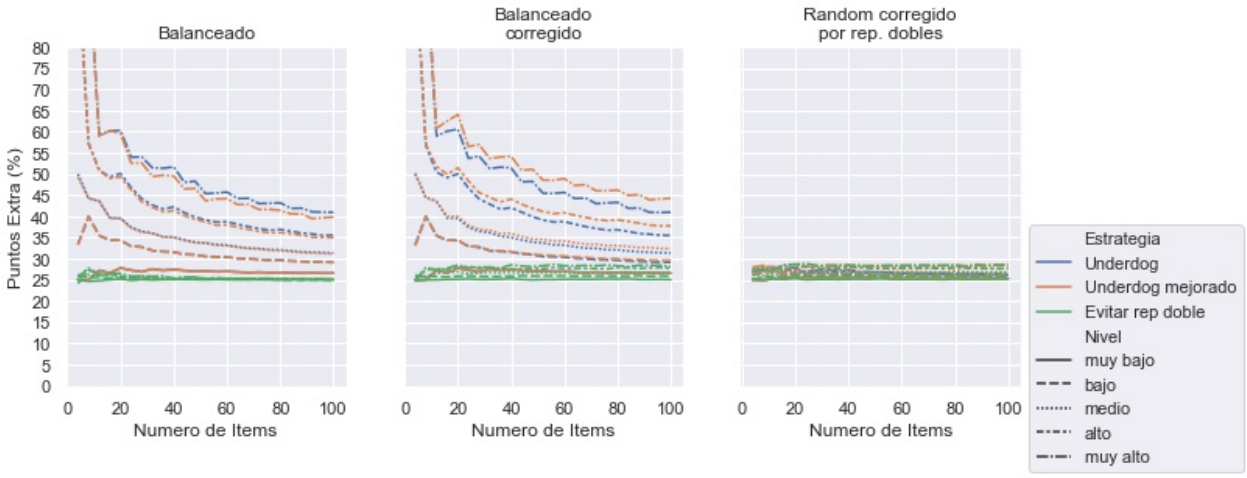
En la Figura 3.9 se presenta la ganancia porcentual para todos los niveles de desempeño posibles. Se confirman las observaciones cualitativas de la ganancia porcentual en todos los niveles: 1) en los modelos Balanceado Exacto y Balanceado Corregido la ganancia porcentual de las estrategias Underdog y Underdog Mejorado son crecientes con respecto al nivel del examinado (de hecho, crecen exponencialmente) y 2) en los modelos Balanceado Corregido y Random Corregido por repeticiones dobles, la ganancia porcentual de las estrategias Underdog Mejorado y Evitar Repeticiones Dobles es similar en todos los niveles (linealmente creciente pero con una pendiente muy pequeña).

Figura 3.9: Ganancia porcentual para niveles en $\{1, ..., 99\}$.



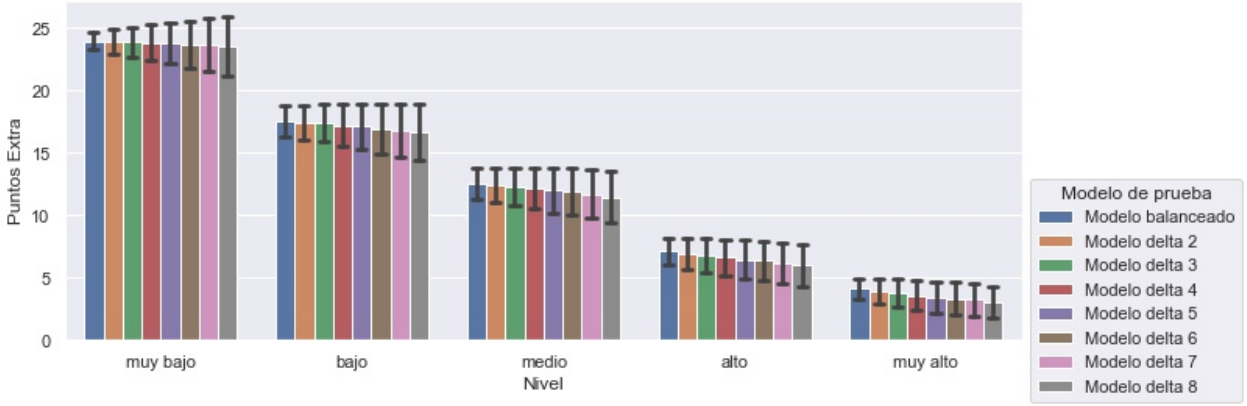
La ganancia porcentual para distinto número de ítems se presenta en la Figura 3.10. Se observa que la ganancia porcentual de las estrategias Underdog y Underdog Mejorado, en los modelos Balanceado y Balanceado Corregido crece drásticamente en pruebas cuyo número de ítems es bajo. Además, se puede notar que la diferencia entre la ganancia porcentual de Underdog Mejorado y Underdog en el modelo Balanceado Corregido crece mientras mayor es el número de ítems. Por otra parte, en los modelos Corregidos (Balanceo Corregido y Random Corregido por repeticiones dobles), el número de ítems no tiene influencia en la ganancia porcentual de la estrategia Evitar Repeticiones Dobles.

Figura 3.10: Ganancia porcentual para número de ítems en $\{4, 8, ..., 100\}$



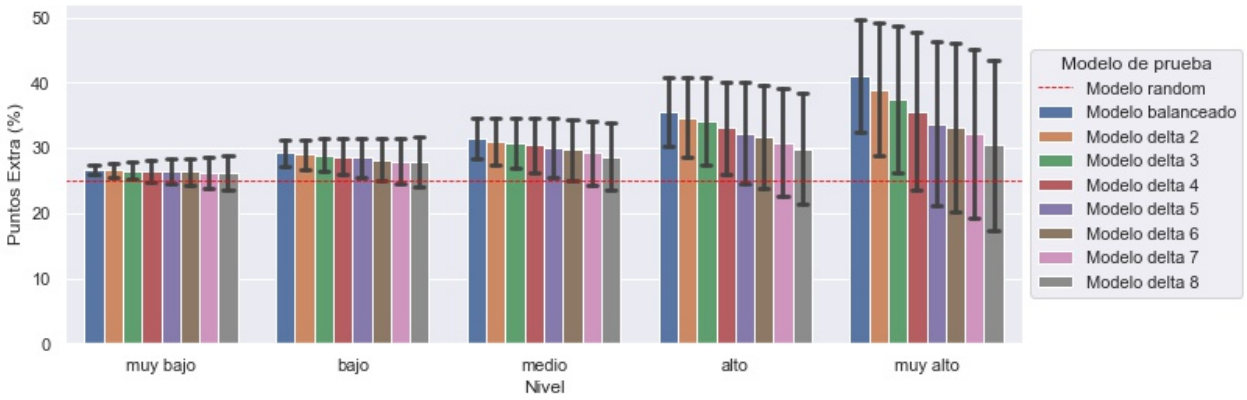
La ganancia de puntos extra obtenidos en cada nivel en los modelos Balanceado Exacto, Balanceado Aproximativo ($\delta \in \{2, ..., 8\}$) y Random con las estrategias Random y Underdog se presentan en las Figuras 3.11 y 3.12, respectivamente

Figura 3.11: Ganancias obtenidas por Underdog para los modelos Balanceado Aproximativo.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 3.12: Ganancias porcentuales obtenidas por Underdog para los modelos Balanceado Aproximativo.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

En todos los modelos excepto en Random, la ganancia de Underdog fue significativamente mayor que la ganancia de la estrategia Random (Balanceado Exacto: Δ_{10}=1.5, d_{10}=0.5; Δ_{40}=2.5, d_{40}=1.0; Δ_{60}=2.6, d_{60}=1.2; Δ_{80}=2.2, d_{80}=1.4; Δ_{90}=1.7, d_{90}=1.5; p-values $<$ 0.0001; Balanceado Aproximativo delta 8: Δ_{10}=1.0, d_{10}=0.3; Δ_{40}=1.6, d_{40}=0.5; Δ_{60}=1.5, d_{60}=0.6; Δ_{80}=0.9, d_{80}=0.5; Δ_{90}=0.5, d_{90}=0.4; p-values $<$ 0.0001). Lo observado indica que, incluso cuando la directriz del balanceo fue relajada a un balanceo aproximativo, Underdog sacó una ventaja (con tamaño desde pequeño hasta mediano) del posicionamiento de claves.

La ganancia de la estrategia Underdog en el modelo Balanceado Exacto fue significativamente mayor o no significativamente diferente a la ganancia en los modelos Balanceado Aproximativo. Además, la ganancia de Underdog fue decreciente con respecto a δ en los modelos balanceo Aproximativo δ, i.e., las ganancias de dos modelos Balanceado Aproximativo, digamos, modelo δ_i y modelo δ_j, con $\delta_i < \delta_j$, o bien fueron similares (p-values $>$ 0.01), o bien la ganancia del modelo δ_i fue significativamente mayor al modelo δ_j. Estos resultados indican que la ganancia que puede sacar Underdog a los modelos de balanceo es cada vez menor en cuanto menos restrictiva sea la recomendación.

Se observa en la Figura 3.12 que la ganancia porcentual de la estrategia Underdog en todos los modelos excepto el modelo Random es creciente con respecto al nivel.

En la Figura 3.13, se presenta la ganancia porcentual en todos los niveles de desempeño posibles. Confirmando la observación de la ganancia porcentual en 5 niveles, se puede ver que la ganancia porcentual de Underdog es creciente con respecto al nivel en todos los modelos excepto el modelo Random. El crecimiento pasa de ser aproximadamente exponencial a aproximadamente lineal a medida que el delta crece.

La ganancia porcentual para distinto número de ítems se presenta en la Figura 3.14. Se observa la ganancia porcentual de la estrategia Underdog aumenta drásticamente en pruebas cuyo número de ítems es bajo para todos los modelos excepto el Random. Además, cuando el número de ítems es bajo, la diferencia entre los distintos modelos decrece.

Figura 3.13: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para niveles en $\{1, ..., 99\}$.

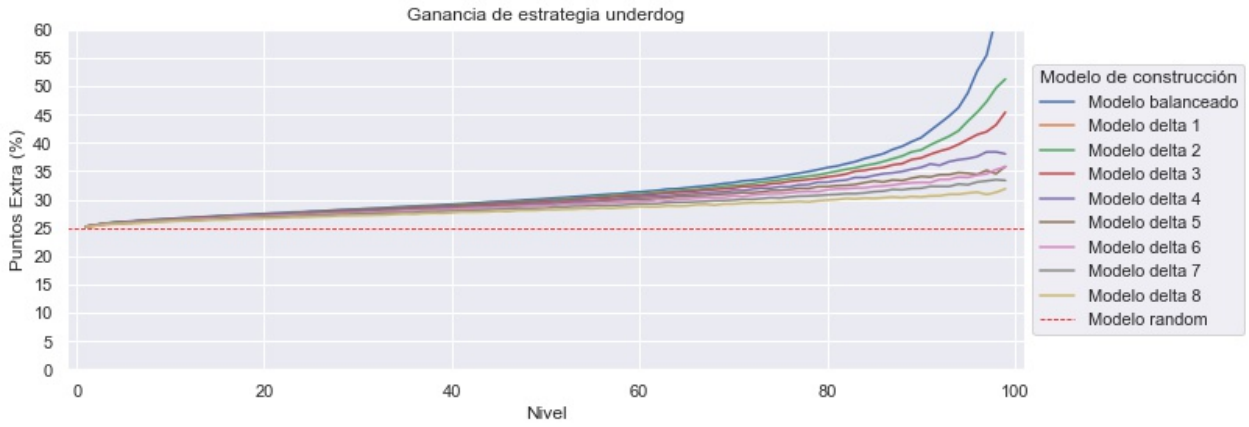
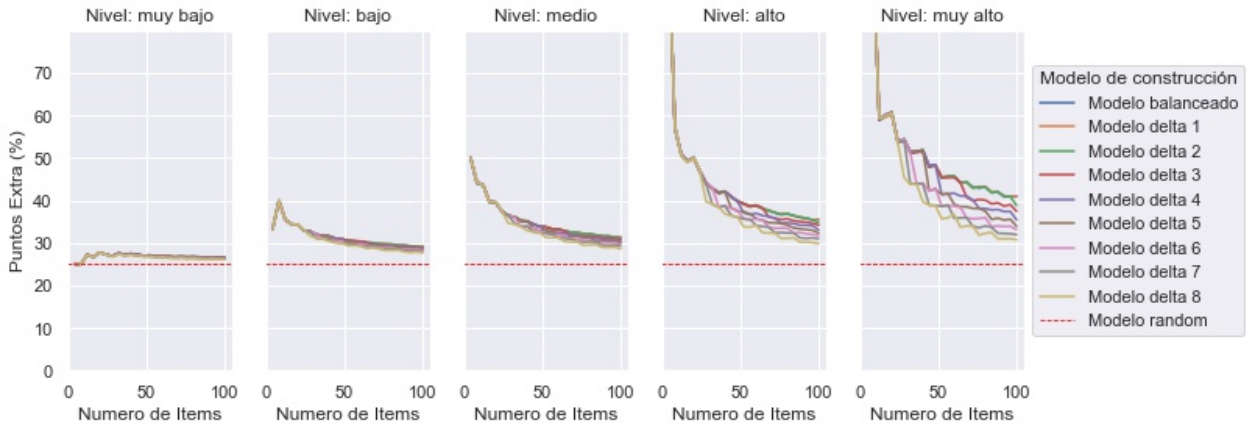


Figura 3.14: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para número de ítems en $\{4, 8, ..., 100\}$



3.3. Discusión

A partir de los resultados del Estudio 2, se puede observar que la ventaja de Underdog reportada en el Estudio 1 se debe esencialmente al balanceo global (al nivel de prueba) y no al balanceo local (en la secuencia de ítems). De hecho, la ganancia de Underdog fue igual en los modelos Balanceado Exacto y Balanceado Corregido, lo que implica que los efectos en el nivel de desempeño y el número de ítems en Balanceado Corregido son debidos a que Underdog en Balanceado Exacto tiene estos efectos. Así, todos los efectos vistos para Balanceado Corregido en el Estudio 1 son a causa de que el Balanceo global es una restricción global que involucra a todos los ítems, y así, mientras mayor información relativa (mayor nivel de desempeño o menor número de ítems en un nivel de desempeño constante) se tenga, mayor será la ganancia porcentual de la estrategia Underdog.

Los resultados también mostraron que la recomendación de balanceo aproximado dejó

espacio a ventaja de las estrategias como Underdog, por lo que, incluso relajando la recomendación a un balanceo aproximado, los modelos ponen en riesgo la validez de los resultados. Al igual que Balanceo Exacto, la ganancia porcentual en los modelos es mayor mientras mayor sea el nivel de desempeño, por lo que relajar la recomendación tampoco arregla el problema de equidad. Para el número de ítems, al igual que en Balanceado Exacto, la ventaja porcentual fue mayor mientras menor fue el número de ítems. Más aún, la estrategia Underdog saca una ventaja similar en pruebas balanceadas de forma exacta y aproximada cuando las pruebas se componen de un número de ítems inferior a 50. Lo anterior pareciera indicar que el balanceo global, en general (aproximado o exacto), es una práctica de posicionamiento de claves subóptima. Es posible que la idea de balancear aproximativamente venga de que la ventaja es cada vez menor mientras mayor es el delta de aproximación, pero a la vista de los datos, para que esto se cumpla el delta a elegir debería ser más grande que los usualmente usados en pruebas.

Respecto a encontrar una estrategia más efectiva que Underdog, la estrategia Underdog Mejorado obtuvo un poco más de ventaja que Underdog en el modelo Balanceado Corregido. Esto indica que el ruido que encontramos en el Estudio 1 para Balanceado Corregido, podría ser aún un poco mayor, por lo que se abre la puerta a que existan otras estrategias que generen una amenaza mayor a la validez de los resultados en los modelos basados en la recomendación de balanceo. Esto es grave porque no se tiene un límite claro de hasta cuantos puntos extras podrían sacar otras estrategias, un límite claro como en el caso de Balanceo Exacto y su estrategia de ganancia maximal Underdog (Corolario A.12), lo que hace aún menos recomendable el uso de estos modelos de construcción.

La recomendación de corregir por repeticiones dobles también dejó espacio a ventaja, aunque pequeña, por parte de estrategias (Underdog Corregido y Evitar Repeticiones Dobles), por lo que el uso de estrategias también ponen en riesgo la validez de los resultados en este modelo. Otro aspecto a considerar es la estabilidad que tuvo la estrategia evitar repeticiones en los diferentes niveles de desempeño y número de ítems. La estabilidad en los niveles de desempeño sugiere que la ventaja en este modelo, al menos para estrategias basadas en evitar repeticiones dobles, no debería tener problemas de equidad, y la estabilidad en el número de ítems sugiere que esta recomendación podría no ser perjudicial en la construcción de pruebas con número de ítems bajo, al contrario del caso del balanceo.

A pesar de que la ventaja de Evitar Repeticiones Dobles no fue alta en este estudio, fue visible, lo que abre la puerta a que se puedan diseñar estrategias optimizadas que sacarían mayor ventaja del balanceo local. Al igual que en el modelo Balanceado Corregido, no es claro tampoco cual es el límite de ganancia que podría sacar una estrategia. Es necesario, entonces, profundizar el análisis de la ventaja posible de obtener con estrategias de evitación de repeticiones. Considerando que evitar repeticiones dobles pudo obtener una ventaja significativa en un modelo que evita series de tres claves, pero que diferentes recomendaciones de la literatura aconsejan evitar una extensión diferente de series, vale la pena analizar cómo interactúan otros modelos y estrategias de evitación de repeticiones (simples y triples), y será entonces el tema del siguiente estudio.

Capítulo 4

Estudio 3: Ganancia en los modelos de evitación de repeticiones

Con el fin de estudiar la posible ganancia en modelos basados en las diferentes recomendaciones de evitación de repeticiones, tres modelos para la evitación de distintos números de repeticiones fueron estudiados. Fue también construida una estrategia que podría sacar provecho de cada modelo.

4.1. Método

El diseño general del Estudio 3 es el mismo que en el Estudio 1.

4.1.1. Modelos de construcción

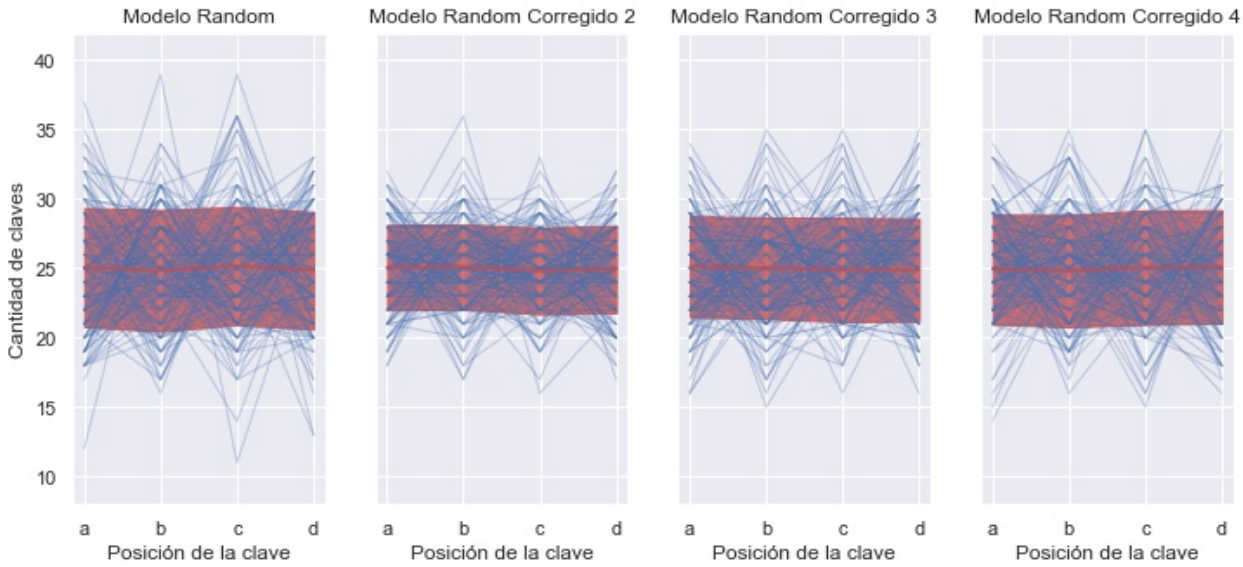
Los tres modelos de construcción basados en la recomendación de evitar repeticiones fueron: Modelo Random Corregido Por Repeticiones Simples, modelo Random Corregido Por Repeticiones Dobles y modelo Random Corregido Por Repeticiones Triples. Cada modelo fue una probabilidad en el conjunto de las pruebas que no tienen el tipo de repeticiones correspondiente (simple, doble, o triple). La probabilidad de máxima entropía en este caso fue la probabilidad uniforme sobre tal conjunto (Corolario A.3).

10.000 simulaciones fueron realizadas para verificar los modelos. Con respecto al número de claves en cada opción, acorde a lo esperado, se observa en la Figura 4.1 que el número de claves tiene un promedio y una varianza similar en cada opción para todos los modelos, coincidiendo con la idea de que evitar repeticiones tiene poca influencia en el número de claves de cada opción.

En las Figuras 4.2 y 4.3, se confirma que el modelo Random Corregido Por Repeticiones Simples no tiene repeticiones simples, y que el modelo Random tiene más repeticiones simples que el modelo Random Corregido Por Repeticiones Triples, y este que el modelo Random

Corregido Por Repeticiones Dobles. En la Figuras 4.4 y 4.5, se confirma que el modelo Random Corregido Por Repeticiones Dobles no tiene repeticiones dobles y que el modelo Random tiene más repeticiones dobles que el modelo Random Corregido Por Repeticiones Triples. En las Figuras 4.6 y 4.7, se confirma que el modelo Random Corregido Por Repeticiones Triples no tiene repeticiones triples y que el modelo Random tiene pocas repeticiones triples. En resumen, los indicadores calculados coinciden con lo que se espera obtener a partir de los supuestos teóricos.

Figura 4.1: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 4.2: Porcentaje de pruebas con repeticiones simples

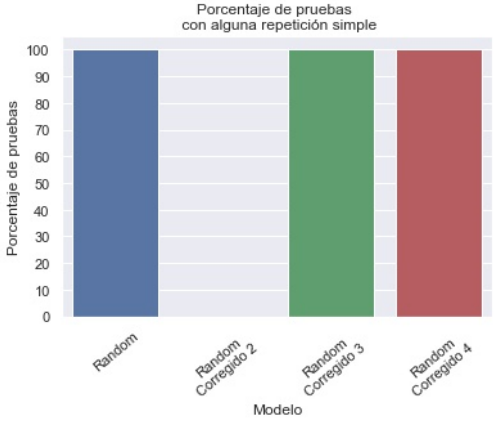
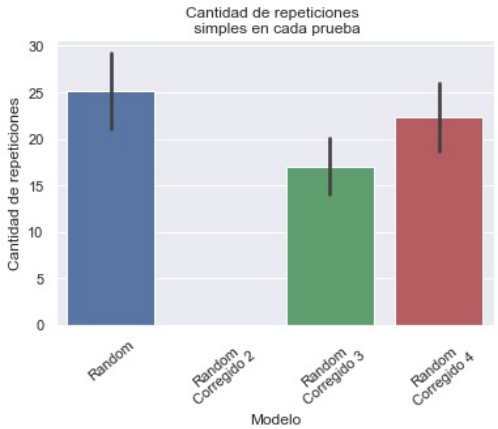


Figura 4.3: Cantidad de repeticiones simples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 4.4: Porcentaje de pruebas con repeticiones dobles

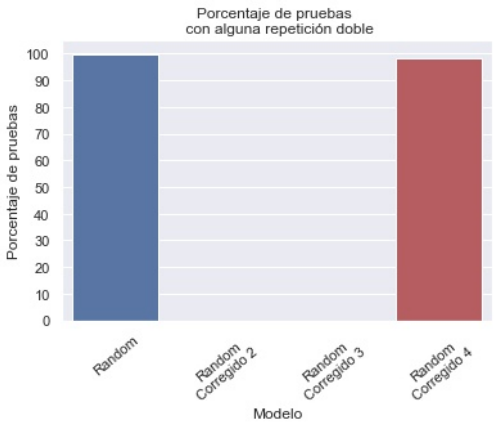
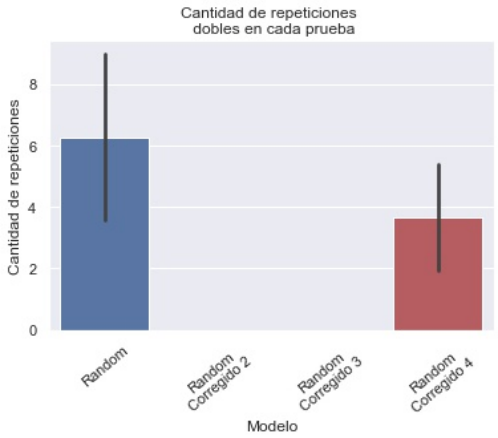


Figura 4.5: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 4.6: Porcentaje de pruebas con repeticiones triple

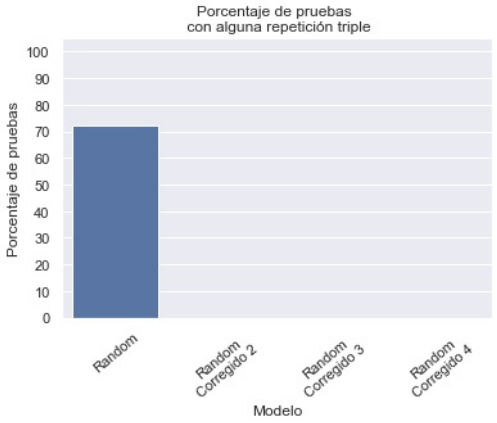
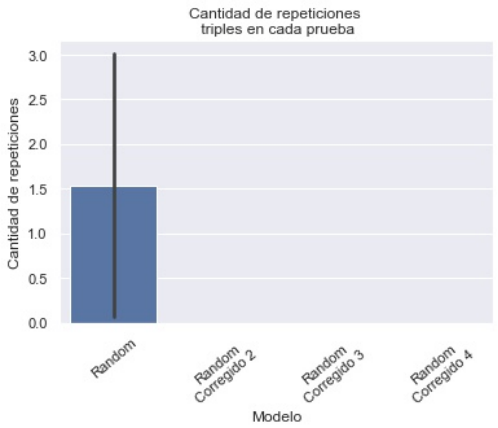


Figura 4.7: Cantidad de repeticiones triples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

4.1.2. Estrategias de resolución

Las tres estrategias simuladas en este estudio fueron Evitar Repeticiones Simples, Evitar Repeticiones Dobles y Evitar Repeticiones Triples. Estas estrategias corresponden a responder los ítems por estrategia, secuencialmente, evitando en cada ítem crear una repetición simple, doble o triple con las respuestas actuales. En caso de que haya más de una opción que no genere una repetición simple, doble o triple, se eligió aleatoria y uniformemente entre los candidatos.

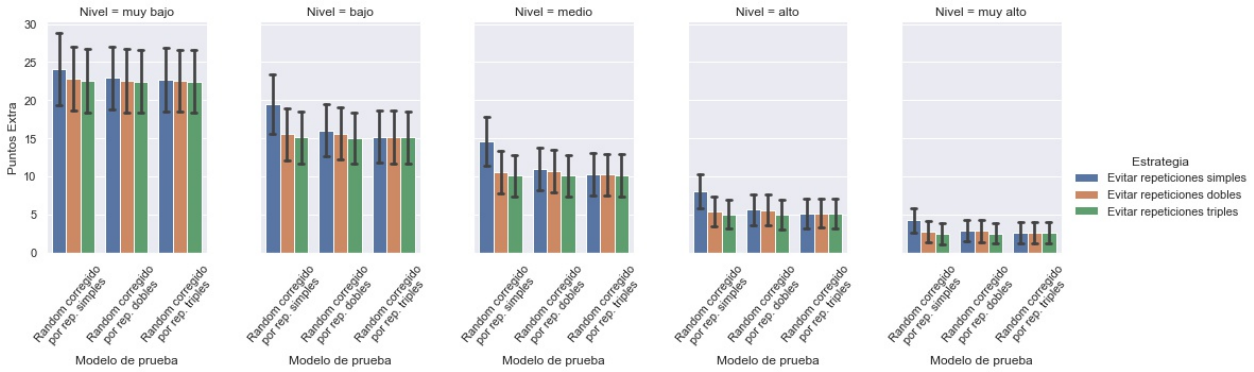
4.1.3. Análisis

Se hicieron los mismos análisis que en el Estudio 1.

4.2. Resultados

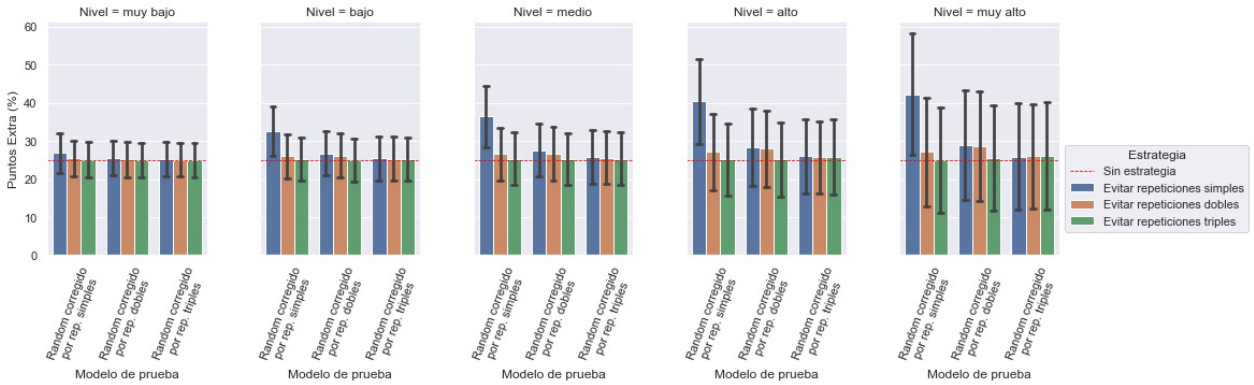
Para los modelos Random, Random Corregido Por Repeticiones Simples, Random Corregido Por Repeticiones Dobles y Random Corregido Por Repeticiones Triples, y para las estrategias Random, Evitar Repeticiones Simples, Evitar Repeticiones Dobles y Evitar Repeticiones Triples, la ganancia y la ganancia porcentual obtenidas en 100 ítems y 5 niveles son presentadas en las Figuras 4.8 y 4.9, respectivamente.

Figura 4.8: Ganancias obtenidas para el Estudio 3.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 4.9: Ganancias porcentuales obtenidas para el Estudio 3.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

En el modelo Random Corregido Por Repeticiones Simples, la ganancia de la estrategia Evitar Repeticiones Simples fue mayor (con tamaños desde pequeños a muy grandes) a la de la estrategia Random (Δ_{10}=1.8, d_{10}=0.4; Δ_{40}=4.4, d_{40}=1.2; Δ_{60}=4.5, d_{60}=1.5; Δ_{80}=3.1, d_{80}=1.5; Δ_{90}=1.8, d_{90}=1.2; p-values < 0.001), la ganancia de la estrategia Evitar Repeticiones

Dobles fue igual o levemente mayor (con efectos despreciables o pequeños) a la de la estrategia Random (Δ_{10}=0.3, d_{10}=0.1; Δ_{40}=0.6, d_{40}=0.2; Δ_{60}=0.6, d_{60}=0.2; Δ_{80}=0.4, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p-values < 0.0001) y la ganancia de la estrategia Evitar Repeticiones Triples fue similar (tamaños de efecto despreciables) a la de la estrategia Random (d_{40}, $d_{90} = 0.0$; p_{10}, p_{60}, $p_{80} > 0.1$). Estos resultados indican que evitar todo tipo de repetición da lugar a que se pueda obtener una ventaja importante (tamaño de efecto mayor a 1.2 para los niveles mayores a 10) gracias al uso de una conducta estratégica adaptada.

En los modelos Random Corregido Por Repeticiones Dobles, la ganancia de la estrategia Evitar Repeticiones Simples fue mayor (tamaños de efecto pequeños) que la ganancia de Random (Δ_{10}=0.4, d_{10}=0.1; Δ_{40}=1.1, d_{40}=0.3; Δ_{60}=1.0, d_{60}=0.4; Δ_{80}=0.6, d_{80}=0.3; Δ_{90}=0.4, d_{90}=0.3; p-values < 0.0001), al igual que la estrategia Evitar Repeticiones Dobles (Δ_{10}=0.1, d_{10}=0.0; Δ_{40}=0.6, d_{40}=0.2; Δ_{60}=0.6, d_{60}=0.2; Δ_{80}=0.6, d_{80}=0.3; Δ_{90}=0.4, d_{90}=0.3; p-values < 0.0001). La ganancia de la estrategia Evitar Repeticiones Triples, por otra parte, fue similar (tamaños de efecto despreciables) a Random (Δ_{40}=0.2; Δ_{60}=0.1; Δ_{80}=0.1; Δ_{90}=0.1, d_{90}=0.1; d_{10}, d_{40}, d_{60}, $d_{80} = 0.0$; p_{40}, p_{60}, p_{80}, $p_{90} < 0.0001$, $p_{10} = 0.5$). Estos resultados muestran que para las repeticiones dobles existen incluso otras estrategías (como Evitar Repeticiones Simples), que sin ser directamente diseñada para sacar provecho de la evitación de repeticiones dobles, tienen un poco más de ganancia que la estrategia Evitar Repeticiones Dobles, mostrada en el Estudio 1. Esto sugiere que no es claro hasta cuantos puntos podría dar una estrategia en este modelo.

En los modelos Random Corregido por repeticiones triples, las ganancias de todas las estrategias fueron similares (tamaños de efecto despreciables) a Random (Evitar rep. simples: d_{10}, d_{40}, d_{60}, d_{80}, $d_{90} = 0.1$; p-values < 0.0001; Evitar rep. dobles: d_{40}, d_{60}, d_{80}, $d_{90} = 0.1$; p_{40}, p_{60}, p_{80}, $p_{90} < 0.0001$, $p_{10} = 0.5$; Evitar rep. triples: d_{60}, d_{80}, $d_{90} < 0.1$; p_{60}, p_{80}, $p_{90} < 0.0001$; p_{10}, $p_{40} > 0.1$), indicando que estas estrategias no obtienen ganancia cuando el modelo sólo evita repeticiones triples.

En la Figura 4.9, se observa que la ganancia porcentual de la estrategia Evitar Repeticiones Simples en el modelo Random Corregido Por Repeticiones Simples es creciente con respecto al nivel. No obstante, para todos los demás casos la ganancia es similar en todos los niveles.

En la Figura 4.10, se presenta la ganancia porcentual para todos los niveles de desempeño posibles. Confirmando las observaciónes de la ganancia porcentual en 5 niveles, se puede ver que la ganancia porcentual de Evitar Repeticiones Simples crece linealmente con una pendiente alta, en el modelo Random Corregido Por Repeticiones Simples, mientras que las otras ganancias crecen linealmente con una pendiente más bien baja.

La ganancia porcentual para distinto número de ítems se presenta en la Figura 4.11. Se observa que en todos los casos, la ganancia porcentual no es afectada por el número de ítems.

4.3. Discusión

Los resultados del Estudio 3 indican que, en general, evitar repeticiones no es una buena estrategia de posicionamiento y que, si se usara, al menos debería cuidarse qué tipo de

Figura 4.10: Ganancia porcentual para niveles en $\{1, ..., 99\}$.

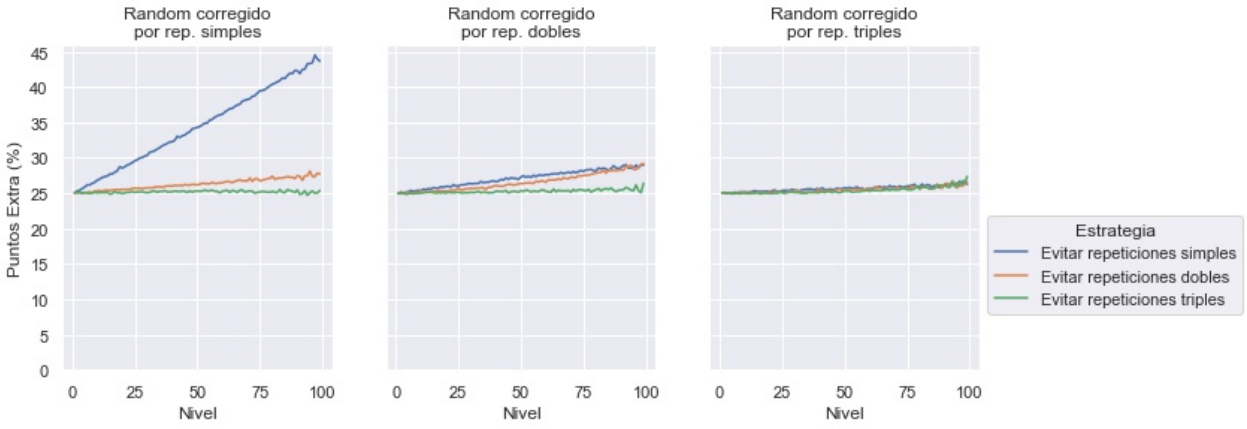
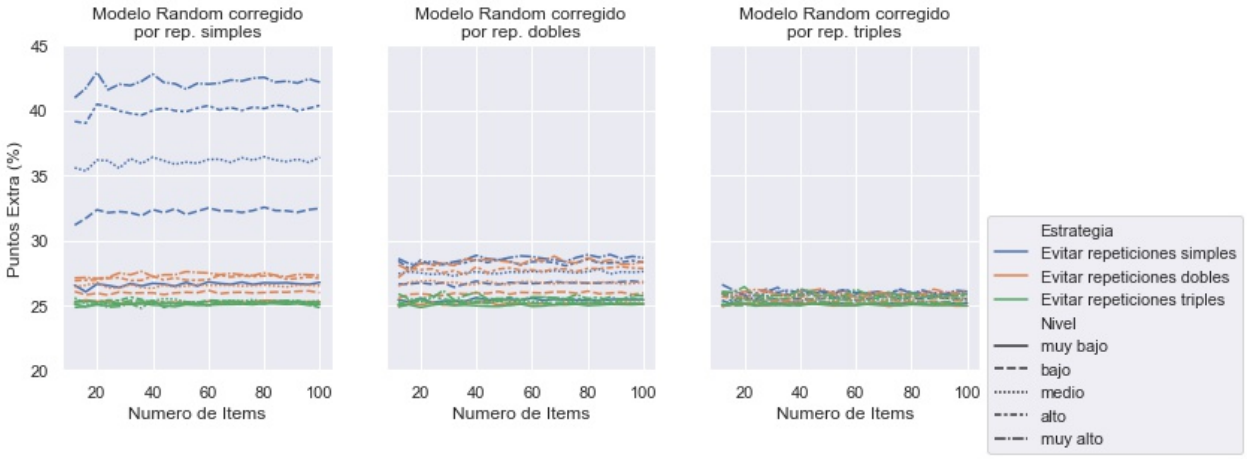


Figura 4.11: Ganancia porcentual para número de ítems en $\{4, 8, ..., 100\}$



repeticiones se evita. Más precisamente, los modelos Random Corregido Por Repeticiones Simples y Random Corregido Por Repeticiones Dobles permitieron ganancias notables con el uso de estrategias basadas en la posición de opciones. Cuando la recomendación seguida fue Evitar Repeticiones Simples, la estrategia Evitar Repeticiones Simples sacó una ventaja incluso comparable a la que sacó la estrategia Underdog en el modelo Balanceo Exacto. Por tanto, Evitar Repeticiones Simples o dobles como recomendación general podría generar problemas de validez en los resultados de una prueba. Al contrario, Evitar Repeticiones Triples tendría poca incidencia en la validez de la prueba.

Para el caso de la estrategia Evitar Repeticiones Simples en el modelo Random Corregido Por Repeticiones Simples, la gran pendiente creciente de la ganancia porcentual con respecto al nivel de desempeño genera inequidades en las mediciones. De una manera más moderada, las estrategias Evitar Repeticiones Simples y Evitar Repeticiones Dobles en el modelo Random Corregido Por Repeticiones Dobles también crean inequidades entre los niveles más bajos y los niveles más altos.

Las diferencias encontradas según número de repeticion podrían deberse a que cuando se evitan repeticiones simples, una respuesta por conocimento en gran medida otorga información sobre las respuestas contiguas porque el conjunto de posibles respuestas baja de 4 a 3, en cambio, cuando se evitan repeticiones dobles se tiene un decrecimiento en la cantidad de respuestas posibles sólo cuando hay una repetición simples en las respuestas por conocimiento, que es menos frecuente. Así, el mismo argumento aplicaría para cuando se ve una repetición triple, lo que sucede mucho menos frecuentemente que ver una repetición doble.

Es notable que la recomendación más frecuente en la evitación de repeticiones sea evitar repeticiones dobles, dado que esta recomendación pone en riesgo la validez de los resultados, y que la única recomendación que parece no amenazar la validez de los resultados (evitar repeticiones triples) sea mucho menos frecuentemente proporcionada.

A lo largo de los Estudios 1, 2 y 3, hemos visto, en pruebas de 4 opciones, que los efectos del uso de estrategias pueden ser perjudicial para la validez de los resultados a una prueba, tanto si no se cuida la posición (sesgo céntrico), como si se cuida con recomendaciones inadecuadas. Aunque cuatro es el número de opciones usado más frecuentemente en pruebas de selección múltiple [25], muchas instituciones importantes, como el DEMRE o el EUNACOM en Chile, usan pruebas de 5 opciones, y existen también constructores que diseñan pruebas de 3 opciones, siguiendo una recomendación usual sobre número de opciones [34]. Los efectos del uso de estrategias podrían tener distinta magnitud y estabilidad (tanto como para el nivel de desempeño como para el número de ítems) en pruebas con otro número de opciones, así como posiblemente Evitar Repeticiones Simples tenga una mayor ganancia en 3 opciones por que podría subir la probabilidad de elección al azar desde $\frac{1}{3}$ a $\frac{1}{2}$ (más que en 4 opciones que es desde $\frac{1}{4}$ a $\frac{1}{3}$). Cabe analizar entonces cómo cambian los efectos reportados en los Estudios 1, 2 y 3 cuando se consideran pruebas con 3 o 5 opciones, y será el tema principal del próximo estudio.

Capítulo 5

Estudio 4: Influencia del número de opciones

Este estudio ofrece una replicación de los Estudios 1, 2, y 3, en pruebas de 3 y 5 opciones.

5.1. Método

El diseño de este estudio es el mismo que en los Estudios 1, 2, y 3. A continuación, sólo se describen aquellos modelos, estrategias, y parámetros de análisis que difieren en pruebas de 3 o 5 opciones.

5.1.1. Modelos de construcción

Para verificar las principales características de estos modelos, al igual que en los estudios anteriores, se simularon el número de pruebas calculado en Apéndice B para cada modelo, esta vez para pruebas de 99 y 100 ítems de 3 y 5 opciones, respectivamente (99 ítems porque la prueba debe poder ser balanceada, i.e., con un número de ítems que sea divisible por 3).

Los modelos usados para la replicación del Estudio 1 fueron:

- **Modelo Random**: Probabilidad uniforme sobre el conjunto de secuencias $\{a, b, c\}^N$ o $\{a, b, c, d, e\}^N$ para 3 o 5 opciones, respectivamente.
- **Modelo Balanceado Corregido**: Como los datos usados para construir este modelo en el Estudio 1 no eran de un número de opciones específico, también aplican para 3 y 4 opciones. La probabilidad es una uniforme en el conjunto de las pruebas balanceadas que no tienen repeticiones dobles.
- **Modelo Sesgo Céntrico**: Para 3 y 5 opciones, también fueron usados datos de grupos de ítems presentando una distribución de claves reportada como sesgada hacia el centro

por sus autores. Dos grupos de ítems analizados en dos estudios para 3 opciones [28, 29] y seis grupos de ítems analizados en tres estudios para 5 opciones [1, 29, 30] fueron usados.

Los porcentajes obtenidos para 3 y para 5 opciones fueron respectivamente:

$$P_a = 33{,}7\,\%, P_b = 45{,}9\,\%, P_c = 20{,}2\,\%$$

$$P_a = 18{,}1\,\%, P_b = 21{,}2\,\%, P_c = 21{,}7\,\%, P_d = 20{,}8\,\%, P_e = 17{,}9\,\%$$

De manera similar al caso de 4 opciones, la probabilidad discreta fue:

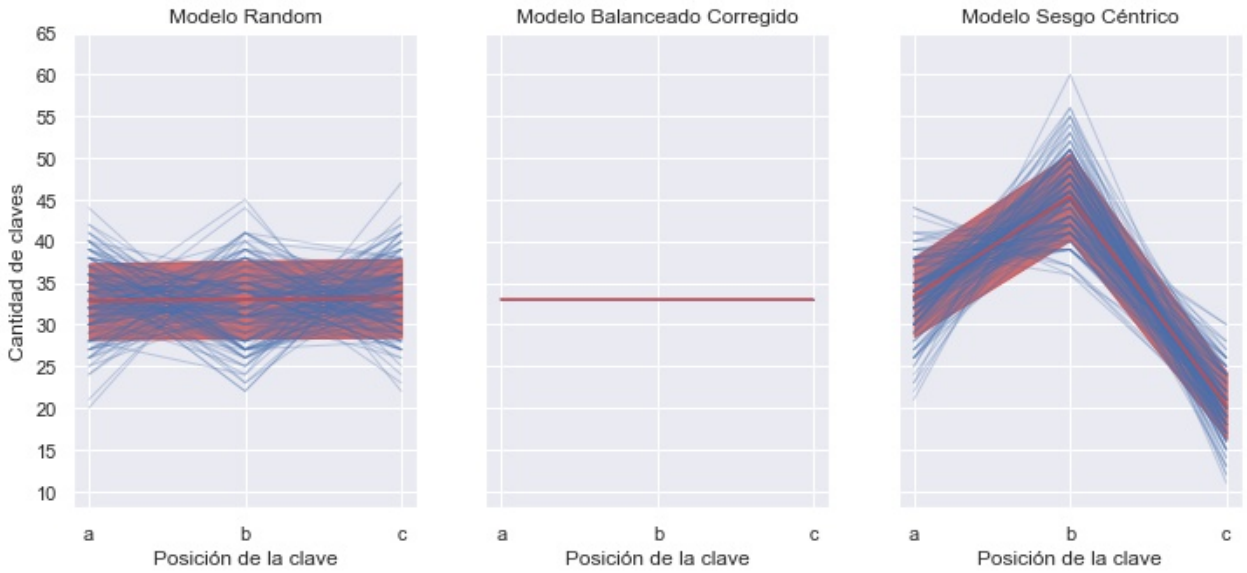
$$\mathbb{P}_{centrica}(\{alt\}) = (P_{alt}/100), \forall alt \in \mathcal{A}$$

Con $\mathcal{A} = \{a, b, c\}$ en 3 opciones y $\mathcal{A} = \{a, b, c, d, e\}$ en 5 opciones. Finalmente, si llamamos $p = (p_1, p_2, ..., p_N)$ una prueba cualquiera, tenemos que su propabilidad de ser generada con el modelo Sesgo Céntrico es:

$$\mathbb{P}_{sesgo\text{-}centrico}(p) = \prod_{i=1}^{N} \mathbb{P}_{centrica}(p_i)$$

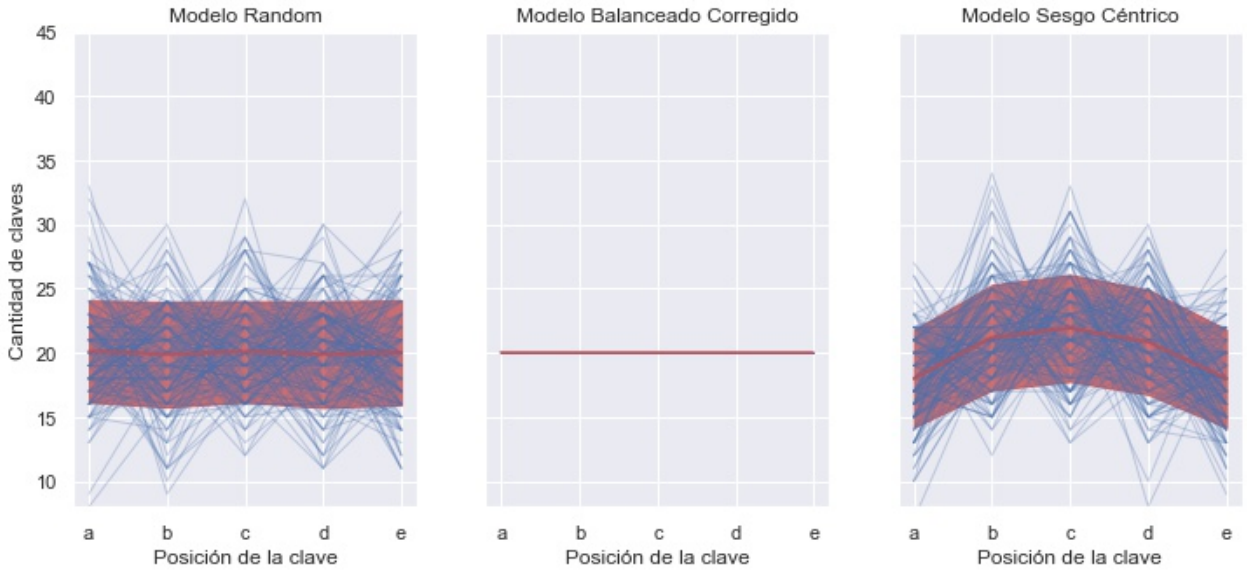
Los indicadores simulados para los modelos del Estudio 1 para 3 y 5 opciones se presentan en las Figuras 5.1, 5.2, 5.3, 5.5, 5.4 y 5.6. Los modelos construidos en este estudio de replicación presentan las características correspondientes a lo esperado teoricamente: 1) el modelo Balanceado Corregido obtiene distribuciones de claves que son perfectamente balanceadas entre opciones, 2) el modelo Random obtiene, en promedio, el mismo número de claves en todas las opciones, pero con una alta varianza, 3) el modelo Sesgo Céntrico tiene una alta varianza y un mayor número de claves en opciones con posiciones céntricas y 4) el modelo Balanceado Corregido no tiene repeticiones dobles, los otros modelos, sí.

Figura 5.1: Estudio 1 en 3 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.2: Estudio 1 en 5 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.3: Estudio 1 en 3 opciones: Porcentaje de pruebas con repeticiones dobles

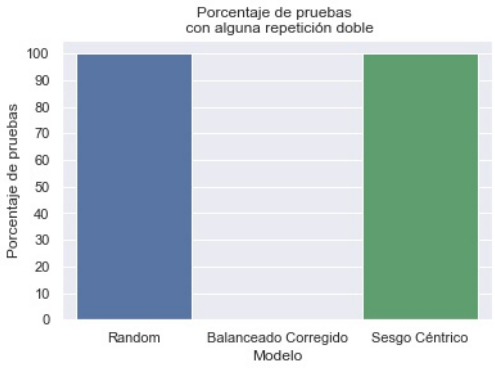
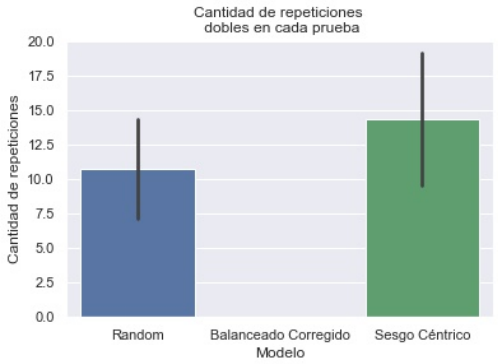


Figura 5.4: Estudio 1 en 3 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.5: Estudio 1 en 5 opciones: Porcentaje de pruebas con repeticiones dobles

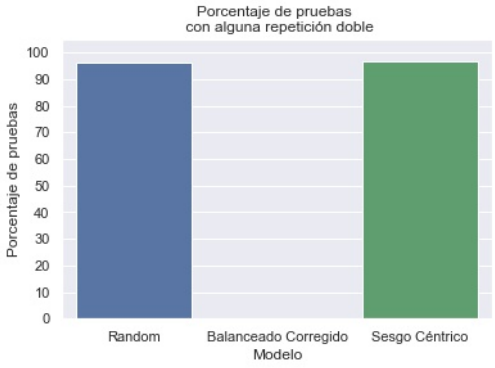
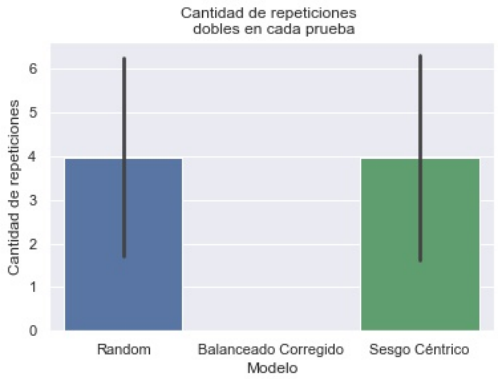


Figura 5.6: Estudio 1 en 5 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



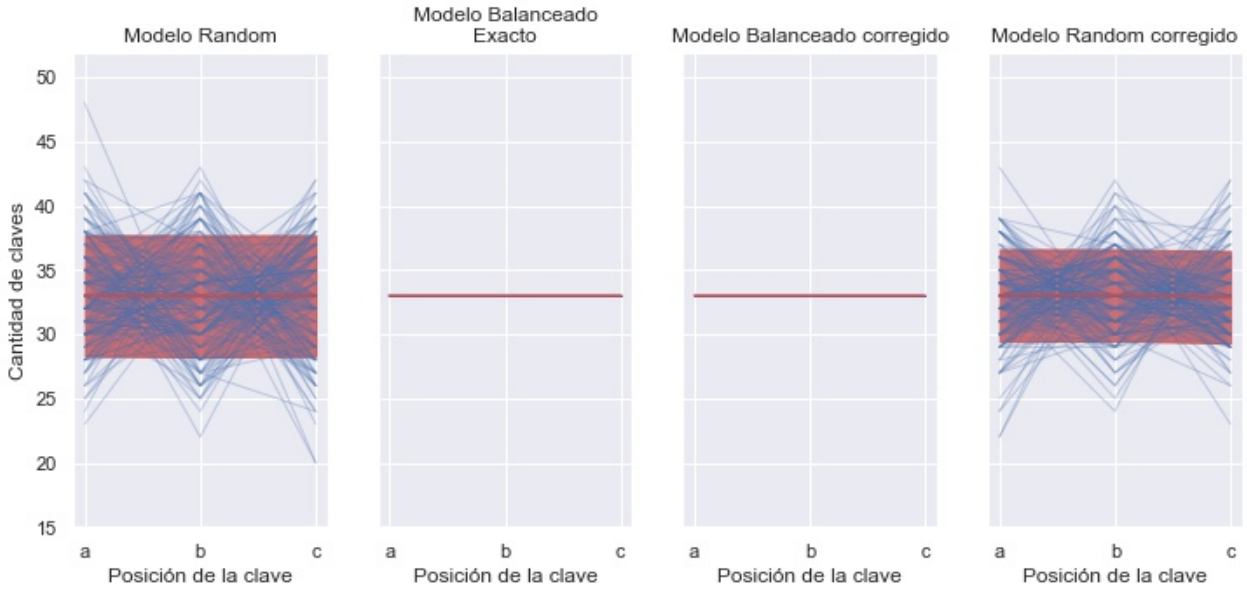
Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Para la replicación del Estudio 2, los modelos fueron:

- **Modelo Balanceado Corregido**: Definido en el Estudio 1.
- **Modelo Random Corregido Por Repeticiones Dobles**: Misma definicion que en el Estudio 2.
- **Modelo Balanceado Exacto**: Misma definicion que en el Estudio 2.
- **Modelo Balanceado Aproximativo**: Como en el Estudio 2, el modelo fue definido usando el criterio delta: la diferencia máxima de número de claves entre opciones. Para el caso de 5 opciones, el criterio delta fue calculado para 12 conjuntos de ítems declarados aproximadamente balanceados por los autores de 4 estudios [1, 36, 38, 9]. Los delta calculados fueron en promedio 4.5 ($\pm$3,8), con un mínimo de 0 y un máximo de 13.3. Como sabemos que delta 0 representa pruebas exactamente balanceadas, consideraremos para 5 opciones delta en $\{1, ..., 14\}$. Como antes, un modelo Balanceado Aproximativo será la probabilidad uniforme en el conjunto de las pruebas cuyo criterio delta sea menor o igual a δ. Para el caso de 3 opciones, no se encontraron datos de pruebas balanceadas aproximadamente, por lo que se consideraron deltas en el rango de los deltas de 4 y 5 opciones, i.e., delta en $\{1, ..., 14\}$.

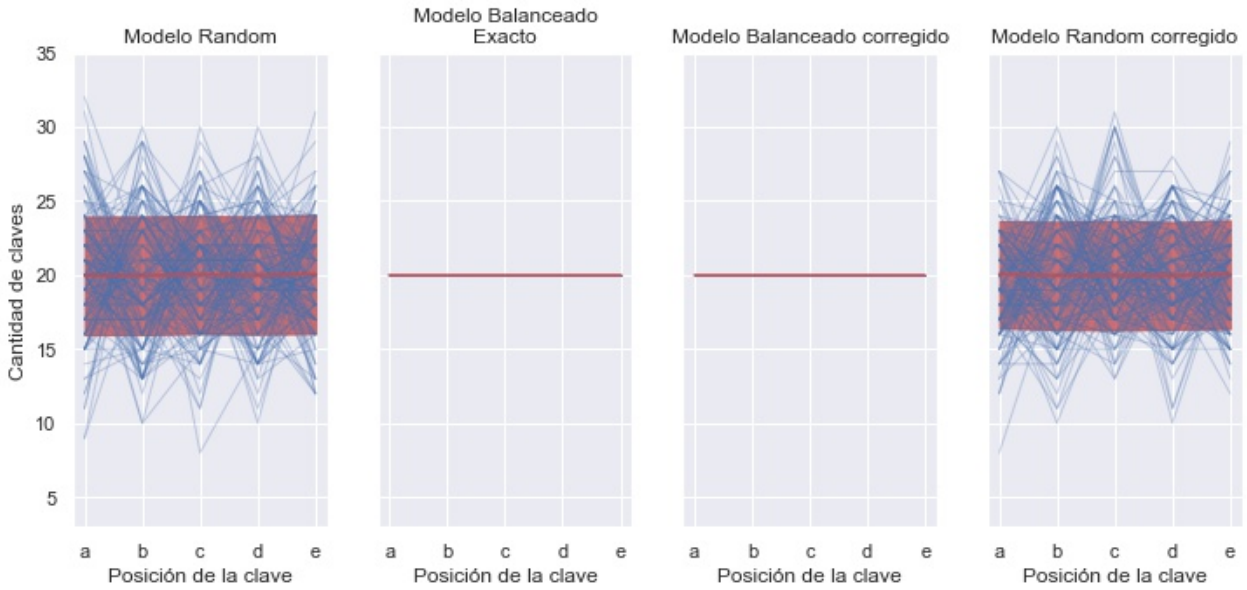
Los indicadores calculados para los modelos del estudio de replicación del Estudio 2 para 3 y 5 opciones se presentan en las Figuras 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17 y 5.18.

Figura 5.7: Estudio 2 en 3 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.8: Estudio 2 en 5 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.9: Estudio 2 en 3 opciones: Porcentaje de pruebas con repeticiones dobles

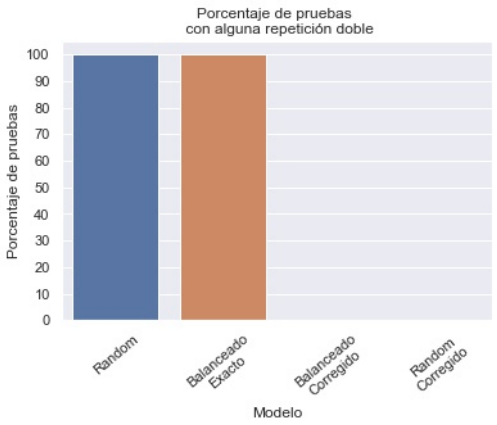
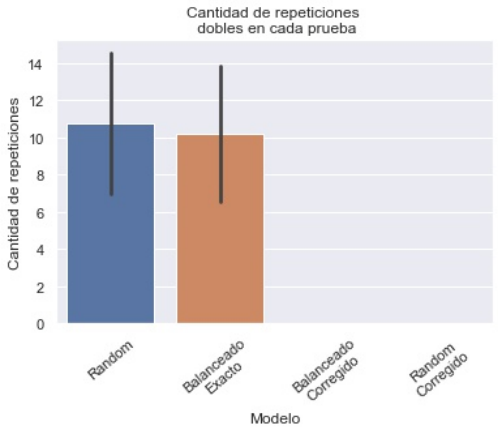


Figura 5.10: Estudio 2 en 3 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.11: Estudio 2 en 5 opciones: Porcentaje de pruebas con repeticiones dobles

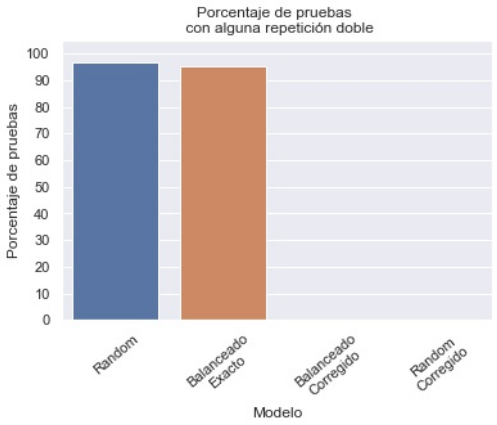
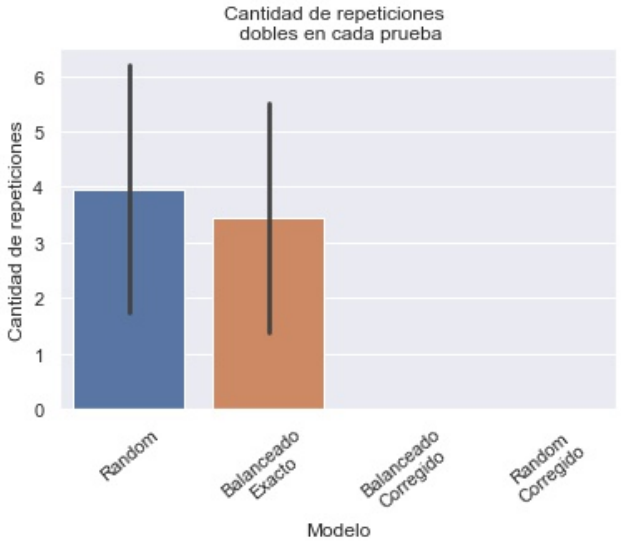
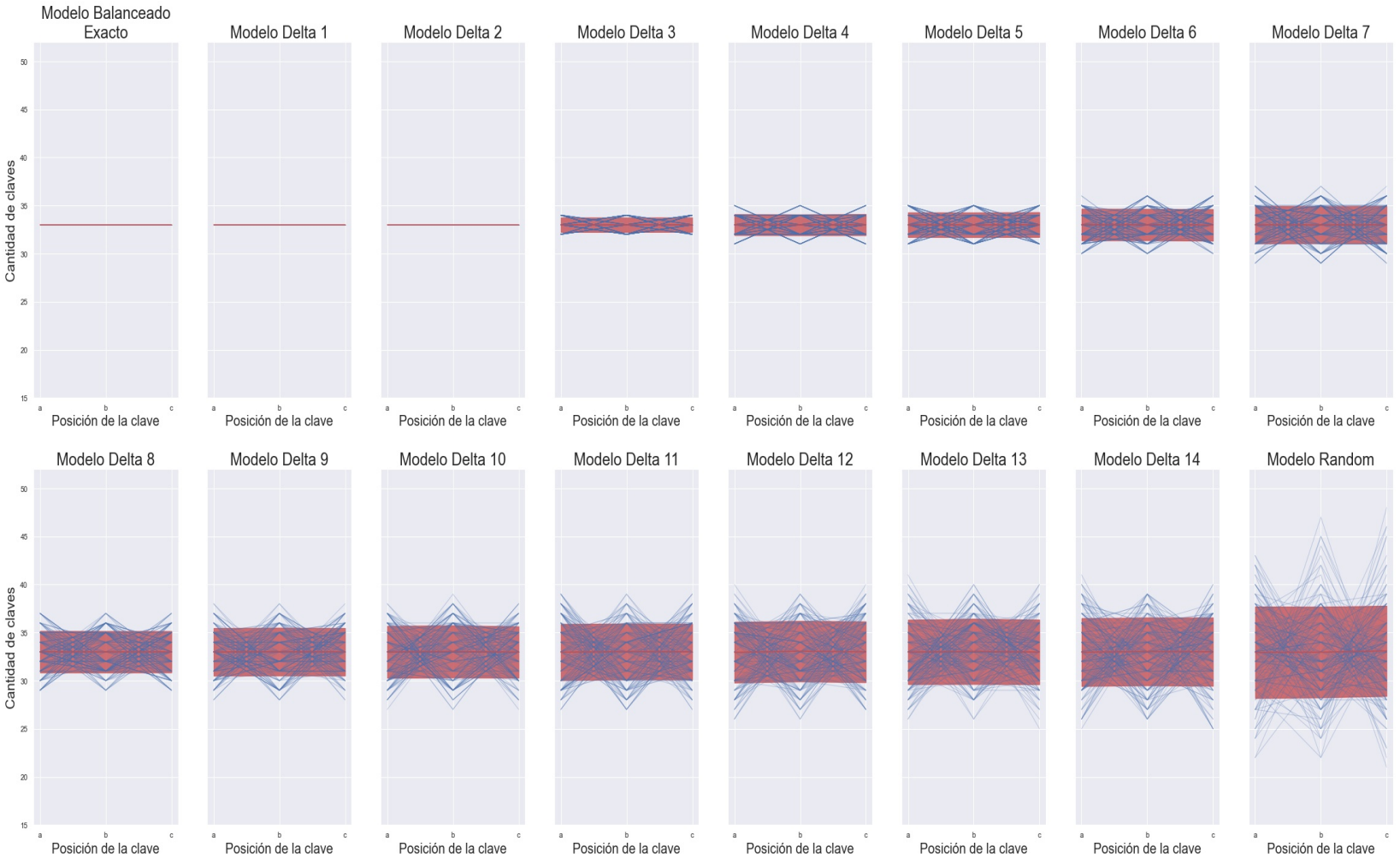


Figura 5.12: Estudio 2 en 5 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



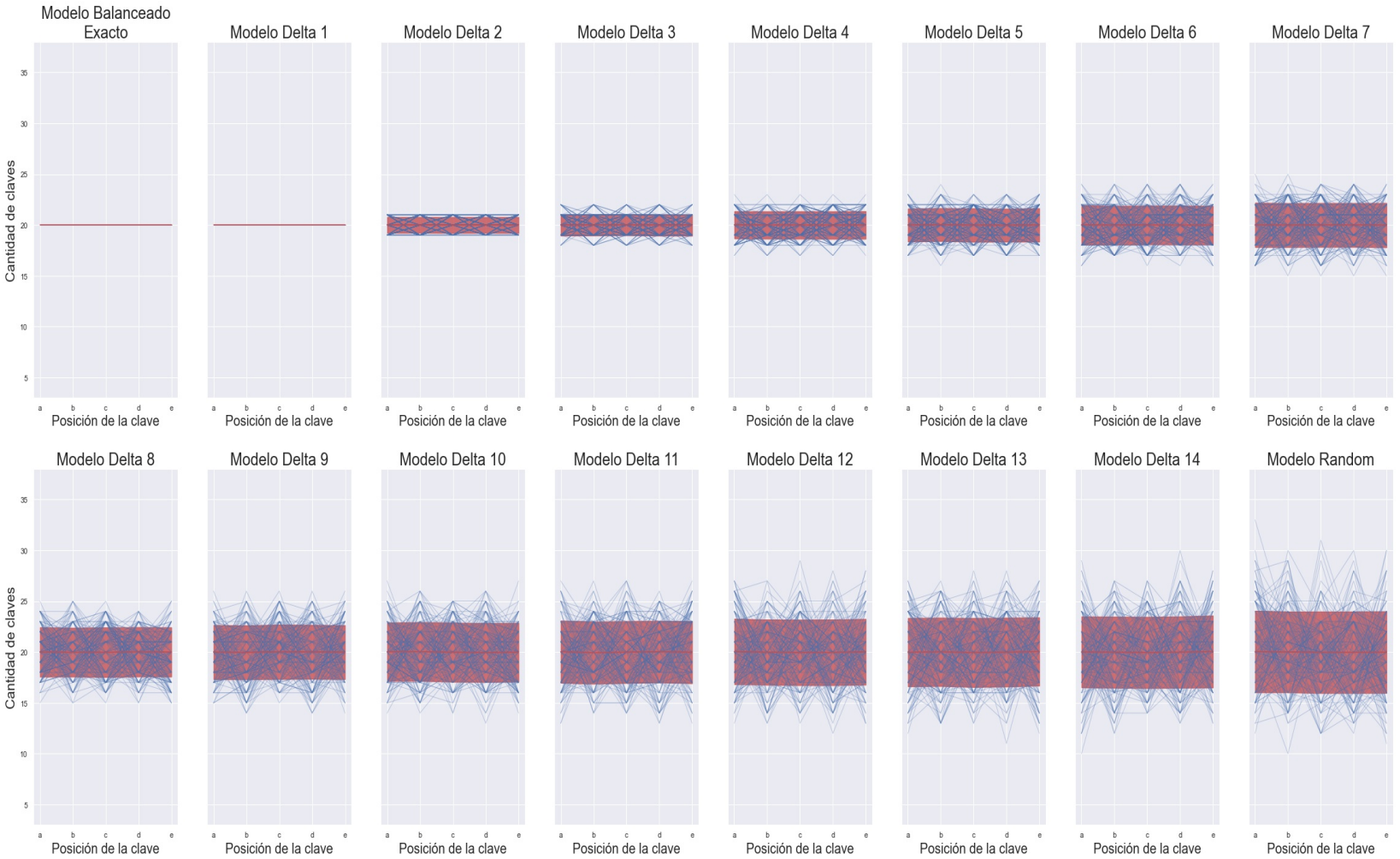
Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.13: Estudio 2 en 3 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.14: Estudio 2 en 5 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.15: Estudio 2 en 3 opciones: Porcentaje de pruebas con repeticiones dobles

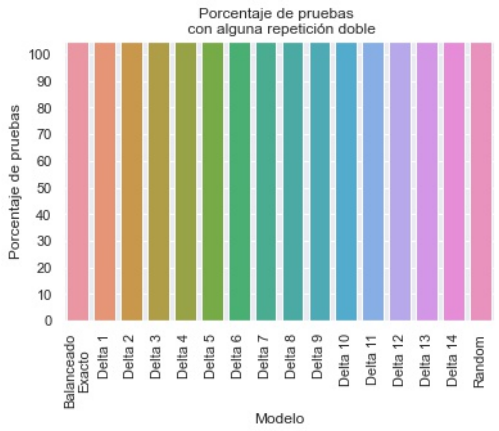
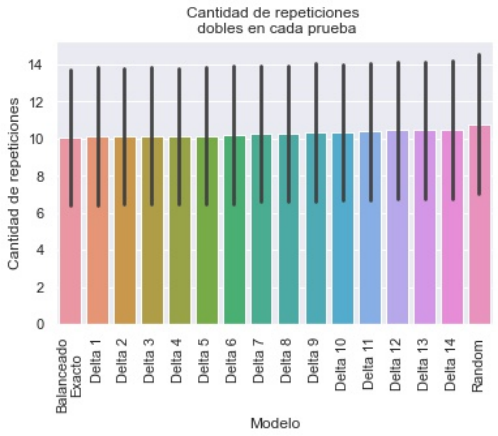


Figura 5.16: Estudio 2 en 3 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.17: Estudio 2 en 5 opciones: Porcentaje de pruebas con repeticiones dobles

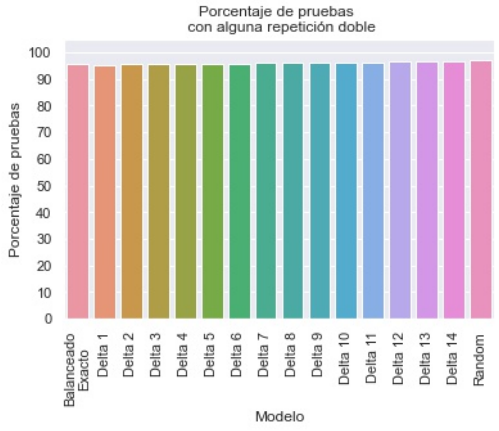
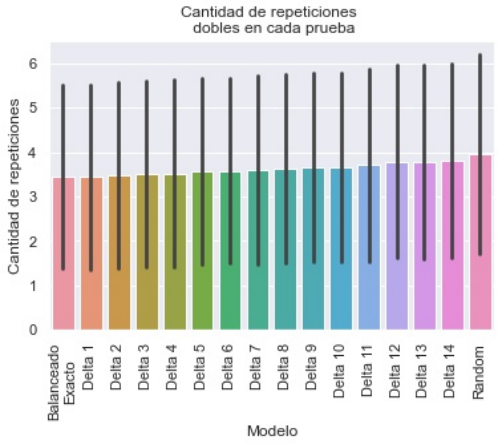


Figura 5.18: Estudio 2 en 5 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.

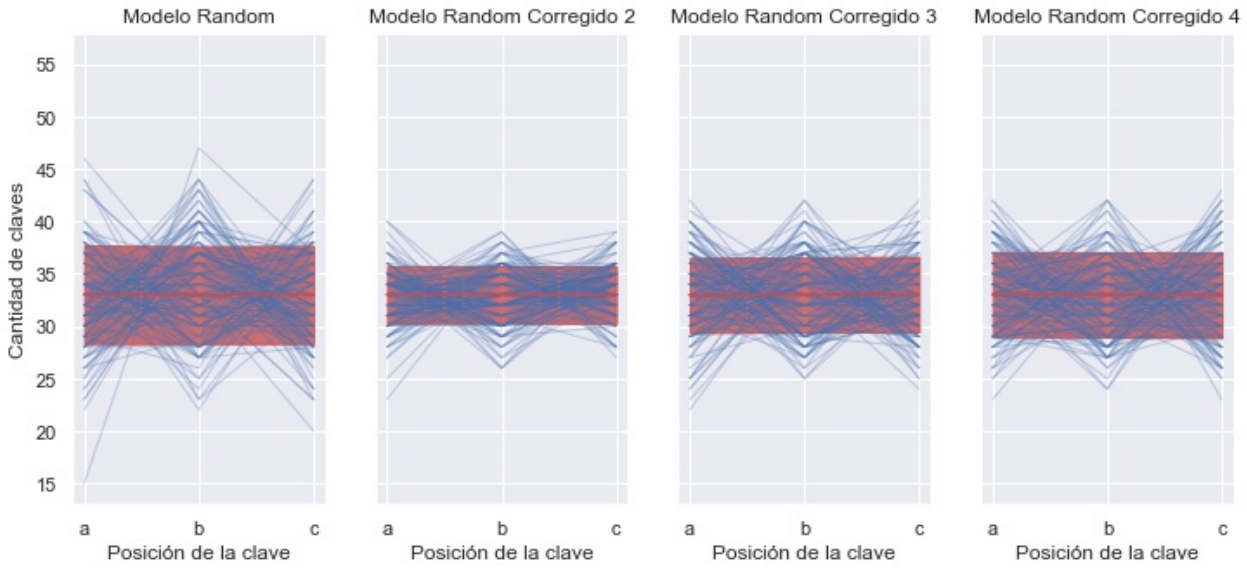


Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

De nuevo, los indicadores calculados se comportan acorde a lo esperado teoricamente: 1) los modelos Random y Random Corregido por repeticiones dobles tienen varianza similar en el número de claves por opción, 2) los modelos Balanceado y Balanceado Corregido tienen el mismo número de claves en todas las opciones, sin varianza y 3) los modelos Corregidos no tienen repeticiones dobles contrario a los modelos no Corregidos. Con respecto a los modelos delta: 1) el número de claves muestra una varianza creciente en delta y las varianzas en estos modelos son mayores que en el modelo Balanceado Exacto y menores que en el modelo Random, 2) el número y porcentaje de las repeticiones dobles son similares en todos los modelos.

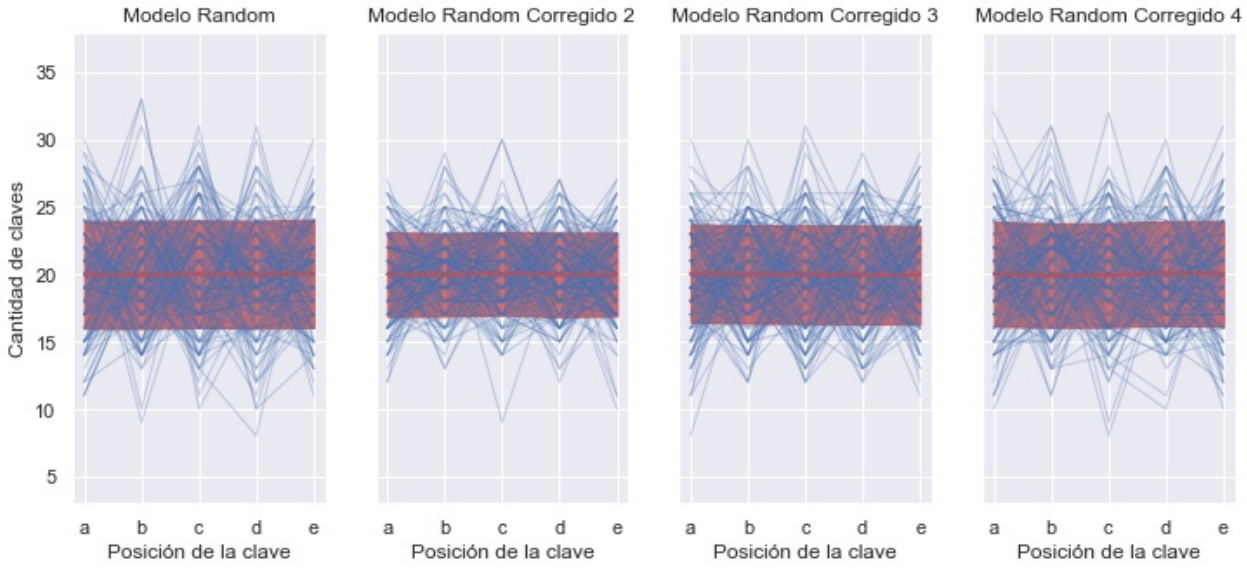
Con respecto al estudio de replicación del Estudio 3, los modelos Random corregido por repeticiones simples, Random corregido por repeticiones dobles y Random corregido por repeticiones triples tienen las mismas definiciones que en el Estudio 3 anterior. Los indicadores nuevamente calculados para pruebas de 3 y 5 opciones se presentan en las Figuras 5.19, 5.20, 5.21, 5.22, 5.23, 5.24, 5.25, 5.26, 5.27, 5.28, 5.29, 5.30, 5.31, 5.32. Nuevamente, los indicadores observados fueron acorde a lo esperado: 1) la distribución del número de claves por opción de los tres modelos es similar, 2) Random Corregido por repeticiones simples no tiene repeticiones simples y Random tiene más repeticiones simples que Random Corregido por repeticiones triples, y este que Random Corregido por repeticiones dobles, 3) Random Corregido por repeticiones dobles no tiene repeticiones dobles y Random tiene más repeticiones dobles que Random Corregido por repeticiones triples y 4) Random Corregido por repeticiones triples no tiene repeticiones triples y Random tiene pocas repeticiones triples.

Figura 5.19: 3 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.20: 5 opciones: Cantidad de claves en cada opción en los diferentes modelos de construcción.



Nota: La línea roja representa la cantidad promedio de claves. La banda roja representa la desviación estándar con respecto al promedio. Cada línea azul corresponde a una de las 150 pruebas escogidas al azar.

Figura 5.21: 3 opciones: Porcentaje de pruebas con repeticiones simples

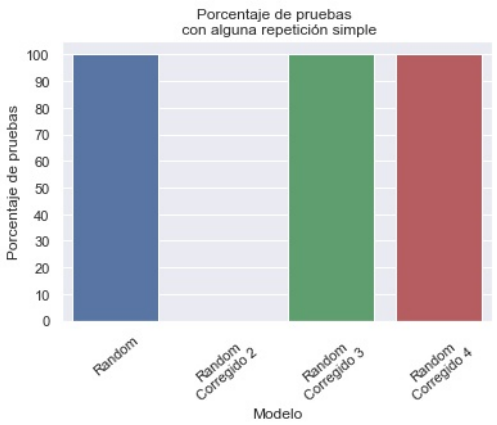
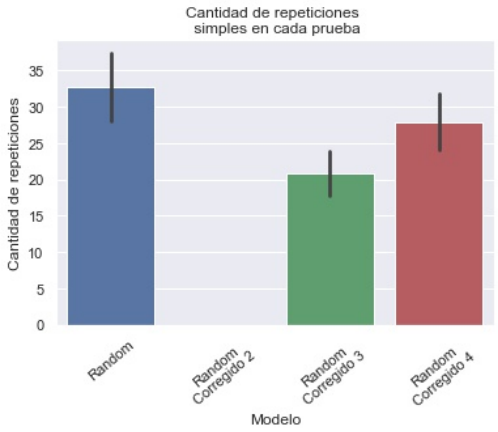


Figura 5.22: 3 opciones: Cantidad de repeticiones simples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.23: 5 opciones: Porcentaje de pruebas con repeticiones simples

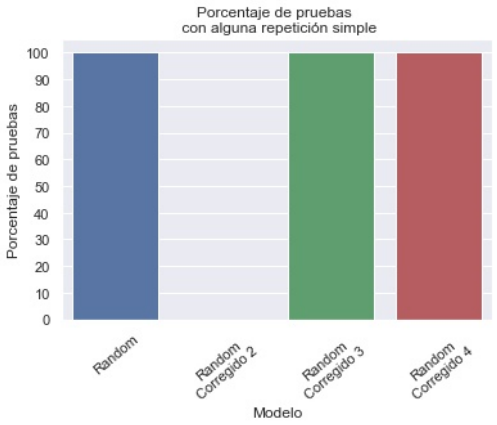
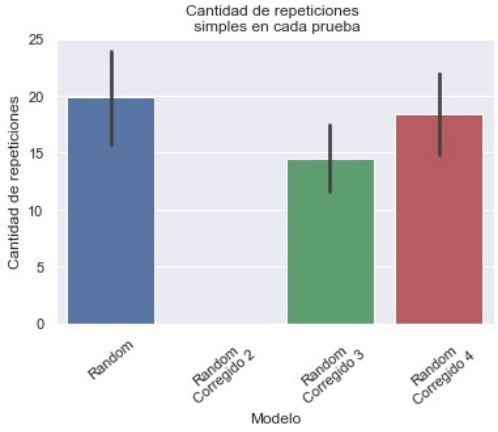


Figura 5.24: 5 opciones: Cantidad de repeticiones simples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.25: 3 opciones: Porcentaje de pruebas con repeticiones dobles

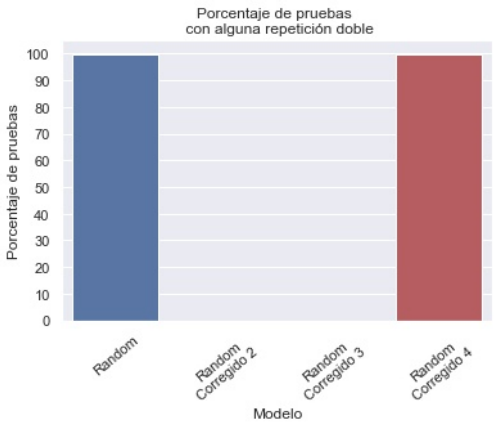
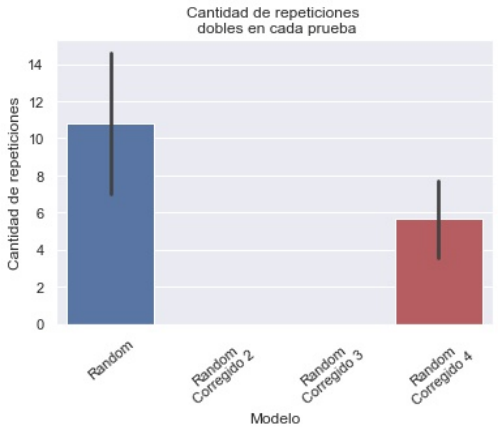


Figura 5.26: 3 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.27: 5 opciones: Porcentaje de pruebas con repeticiones dobles

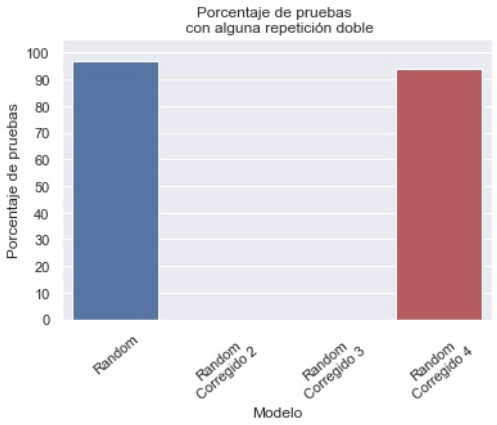
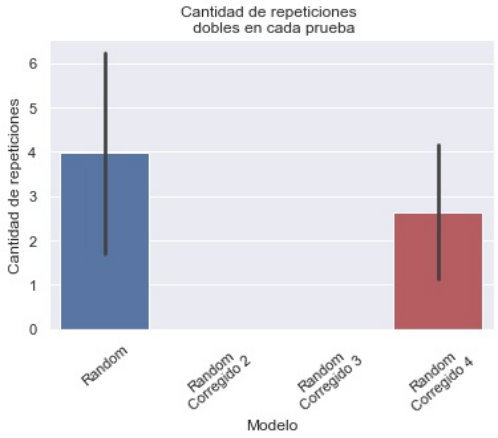


Figura 5.28: 5 opciones: Cantidad de repeticiones dobles de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

5.1.2. Estrategias de resolución

Las estrategias para pruebas de 3 y 5 opciones tendrán las mismas definiciones que para pruebas de 4 opciones, excepto Pura C, que es la única estrategia que depende del número de opciones. Como la idea esencial de la estrategia Pura C es que la clave suele estar en el centro, para 3 opciones quedará definida como Pura B, i.e., la estrategia es rellenar con b las respuestas que faltan. Para el caso de 5 opciones, seguirá siendo Pura C debido a que c sigue

Figura 5.29: 3 opciones: Porcentaje de pruebas con repeticiones triples

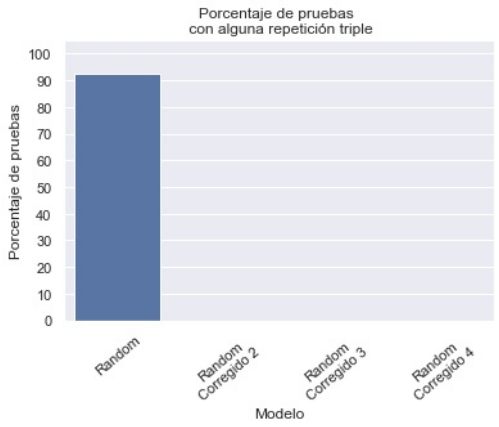
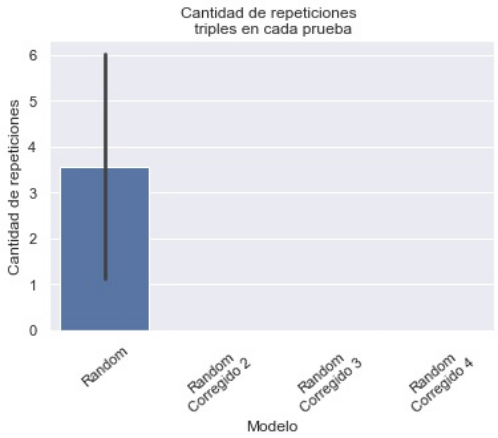


Figura 5.30: 3 opciones: Cantidad de repeticiones triples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.31: 5 opciones: Porcentaje de pruebas con repeticiones triples

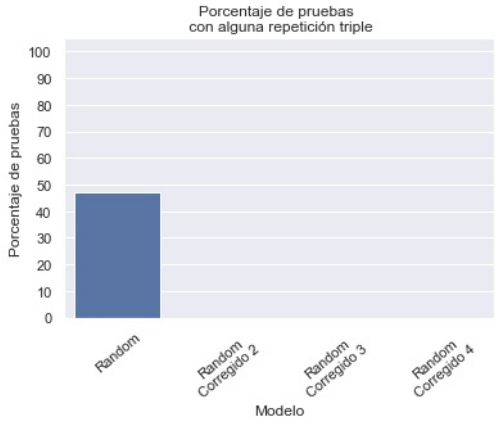
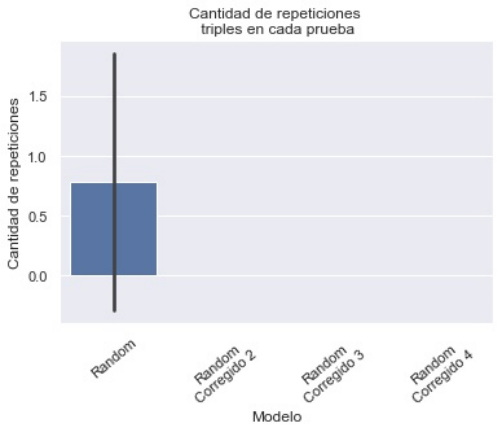


Figura 5.32: 5 opciones: Cantidad de repeticiones triples de cada prueba en los diferentes modelos de construcción.



Nota: Las barras representan la cantidad promedio de repeticiones dobles. Las barras de error representan la desviación estándar con respecto al promedio.

siendo una opción céntrica para este número de opciones.

5.1.3. Análisis

Los análisis de los Estudios 1, 2, y 3 anteriores fueron replicados para pruebas de 3 y 5 opciones. Para los análisis en el número de ítems, los conjuntos fueron $\{3{\cdot}1, 3{\cdot}2, ..., 3{\cdot}33 = 99\}$ para 3 opciones, y $\{5{\cdot}1, 5{\cdot}2, ..., 5{\cdot}20 = 100\}$ para 5 opciones. Para los modelos de construcción

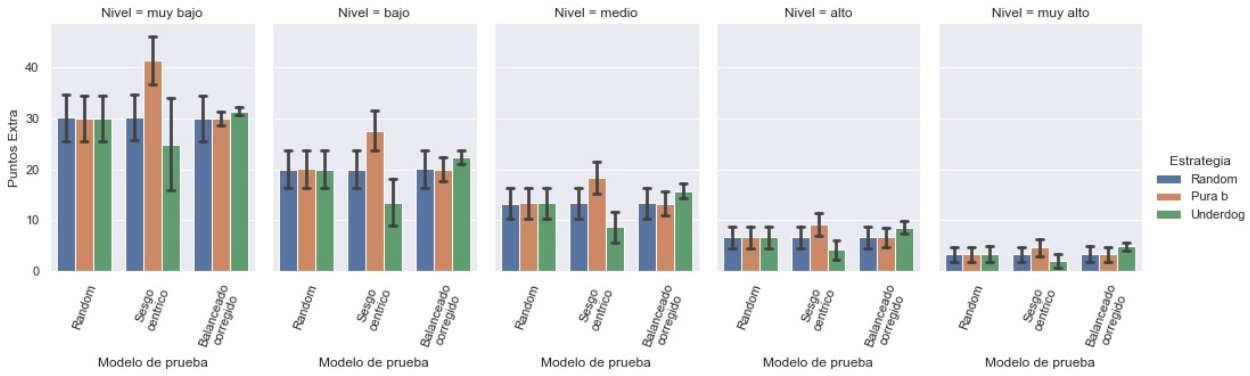
Balanceado Aproximativo se consideraron $\delta \in \{1, ..., 14\}$, para ambos números de opciones.

5.2. Resultados

5.2.1. Replicación del Estudio 1

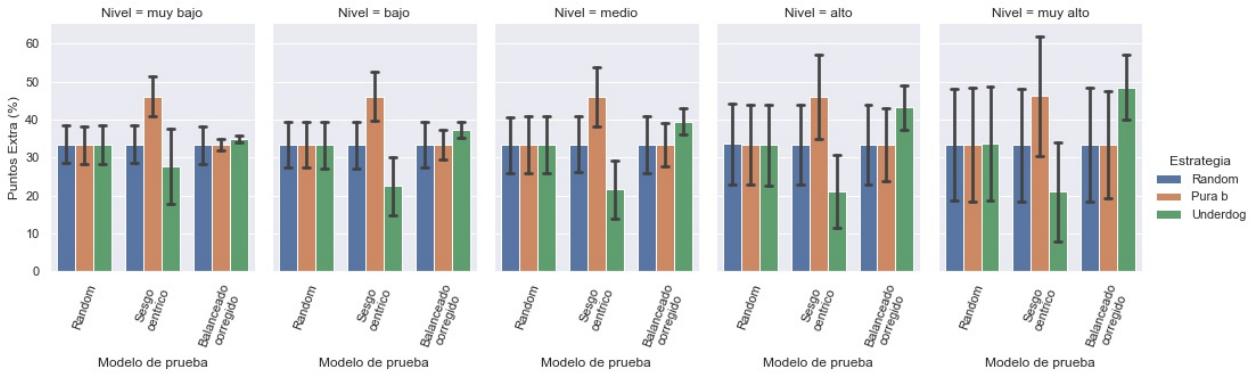
La ganancia y la ganancia porcentual obtenidas en 100 ítems y 5 niveles son presentadas en las Figuras 5.33 y 5.34 para 3 opciones, y en las Figuras 5.35 y 5.36 para 5 opciones.

Figura 5.33: Estudio 1 en 3 opciones: Ganancia para distintos modelos, estrategia y niveles.



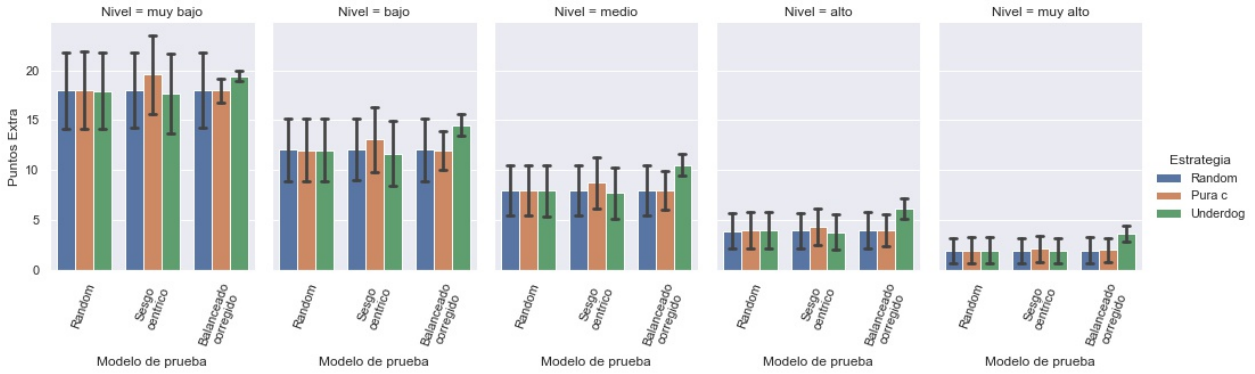
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.34: Estudio 1 en 3 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



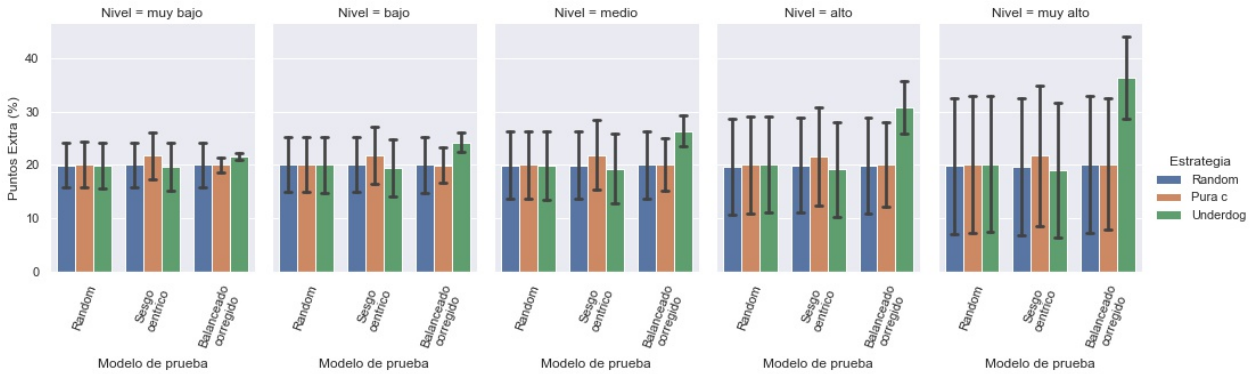
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.35: Estudio 1 en 5 opciones: Ganancia para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.36: Estudio 1 en 5 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Se replican los efectos observados en el Estudio 1 para pruebas de 4 opciones con pruebas de 3 y 5 opciones:

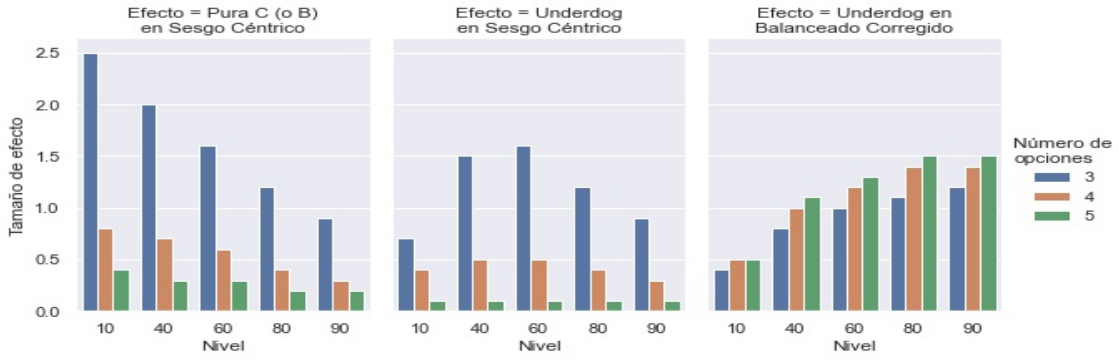
1. En el modelo Sesgo Céntrico, la ganancia de la estrategia Pura C fue significativamente mayor a la ganancia de la estrategia control Random (3 opciones: Δ_{10}=11.3, d_{10}=2.5; Δ_{40}=7.6, d_{40}=2.0; Δ_{60}=5.0, d_{60}=1.6; Δ_{80}=2.5, d_{80}=1.2; Δ_{90}=1.3, d_{90}=0.9; p-values $<$ 0.0001 ; 5 opciones: Δ_{10}=1.6, d_{10}=0.4; Δ_{40}=1.0, d_{40}=0.3; Δ_{60}=0.8, d_{60}=0.3; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p-values $<$ 0.0001) y en el modelo Balanceado Corregido, la ganancia de la estrategia Underdog fue significativamente mayor a la ganancia de la estrategia control Random (3 opciones: Δ_{10}=1.4, d_{10}=0.4; Δ_{40}=2.3, d_{40}=0.8; Δ_{60}=2.4, d_{60}=1.0; Δ_{80}=2.0, d_{80}=1.1; Δ_{90}=1.5, d_{90}=1.2; p-values $<$ 0.0001 ; 5 opciones: Δ_{10}=1.4, d_{10}=0.5; Δ_{40}=2.5, d_{40}=1.1; Δ_{60}=2.5, d_{60}=1.3; Δ_{80}=2.2, d_{80}=1.5; Δ_{90}=1.6, d_{90}=1.5; p-values $<$ 0.0001).
2. En el modelo Random, las ganancias de las tres estrategias no fueron distintas de Random (p-values $>$ 0.1).
3. En el modelo Balanceado Corregido, ni la ganancia de Pura B en 3 opciones ni la

ganancia de Pura C en 5 opciones fue distinta a la ganancia de la estrategia control Random (p-values > 0.1).

En el modelo Sesgo Céntrico, para 3 opciones, al igual que en 4 opciones, la ganancia de la estrategia Underdog no sólo fue significativamente menor a la ganancia de la estrategia Pura B (Δ_{10}=16.5, d_{10}=2.3; Δ_{40}=14.1, d_{40}=3.3; Δ_{60}=9.7, d_{60}=3.1; Δ_{80}=5.0, d_{80}=2.4; Δ_{90}=2.5, d_{90}=1.7; p-values < 0.0001) sino que también fue significativamente menor (con tamaños de efectos desde medianos a muy grandes) a la estrategia control Random (Δ_{10}=5.2, d_{10}=0.7; Δ_{40}=6.4, d_{40}=1.5; Δ_{60}=4.7, d_{60}=1.6; Δ_{80}=2.5, d_{80}=1.2; Δ_{90}=1.2, d_{90}=0.9; p-values < 0.0001). En 5 opciones, por otra parte, la ganancia de Underdog fue menor que la de Pura C (Δ_{10}=1.9, d_{10}=0.5; Δ_{40}=1.4, d_{40}=0.4; Δ_{60}=1.0, d_{60}=0.4; Δ_{80}=0.5, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.2; p-values < 0.0001), pero similar (tamaños de efecto despreciables) a la ganancia de Random (Δ_{10}=0.4; Δ_{40}=0.4; Δ_{60}=0.2; Δ_{80}=0.2; Δ_{90}=0.1; p-values < 0.0001; d_{10},d_{40},d_{60},d_{80},d_{90} = 0.1).

En la Figura 5.37, se puede observar la comparación de los tamaños de efecto encontrados para las pruebas de diferentes números de opciones. Esta comparación indica que las ganancias en el modelo Sesgo Céntrico de las estrategias Pura C (ó B) con respecto a Random bajan mientras mayor sea el número de opciones, mostrando que no preocuparse por el posicionamiento de las claves podría inyectar más ruido en la medición cuando los constructores usan ítems con pocas opciones. Por otro lado, en el modelo Balanceado Corregido, el efecto de la ganancia de la estrategia Underdog con respecto a Random es leve pero consistentemente mayor en 5 opciones que en 4 opciones, y en 4 opciones que en 3 opciones, sugiriendo que los modelos de construcción basados en el balanceo podrían permitir ventajas más fuertes mientras mayor es el número de opciones, en todo caso, en los modelos del Estudio 1 las desventajas de usar un número bajo opciones son claramente mayores a las de usar un número alto de opciones. En el caso la desventaja de Underdog en Sesgo Céntrico, la desventaja es mayor cuanto menor es el número de opciones, de hecho, los tamaños de efecto son similares a los efectos de la ganancia de Pura C (ó B) en Sesgo Céntrico en niveles altos, lo que sugiere que esta desventaja se debe a los porcentajes de la distribución de las claves en el modelo Sesgo Céntrico.

Figura 5.37: Estudio 1: Comparación de los tamaños de efecto encontrados en distinto número de opciones.



Con respecto a las ganancias porcentuales en 3 y 5 opciones, al igual que en 4 opciones, la ganancia porcentual en la estrategia Pura C parece similar en todas los niveles de desempeño,

y la ganancia porcentual de la estrategia Underdog aumenta cuando el nivel aumenta. Con respecto a la puntos porcentuales de desventaja de Underdog en Sesgo Céntrico (que era similar entre los niveles), se replica en 5 opciones que la ganancia porcentual fue similar, pero en 3 opciones esta desventaja muestra una diferencia entre el nivel 10 y los demás niveles, lo que indica que algunos comportamientos podrían ser visibles en un número de opciones pero no en otro.

Las observaciones anteriores se confirman en las Figuras 5.38 y 5.39 para todos los niveles de desempeño posibles, en 3 y 5 opciones respectivamente.

Figura 5.38: Estudio 1 en 3 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.

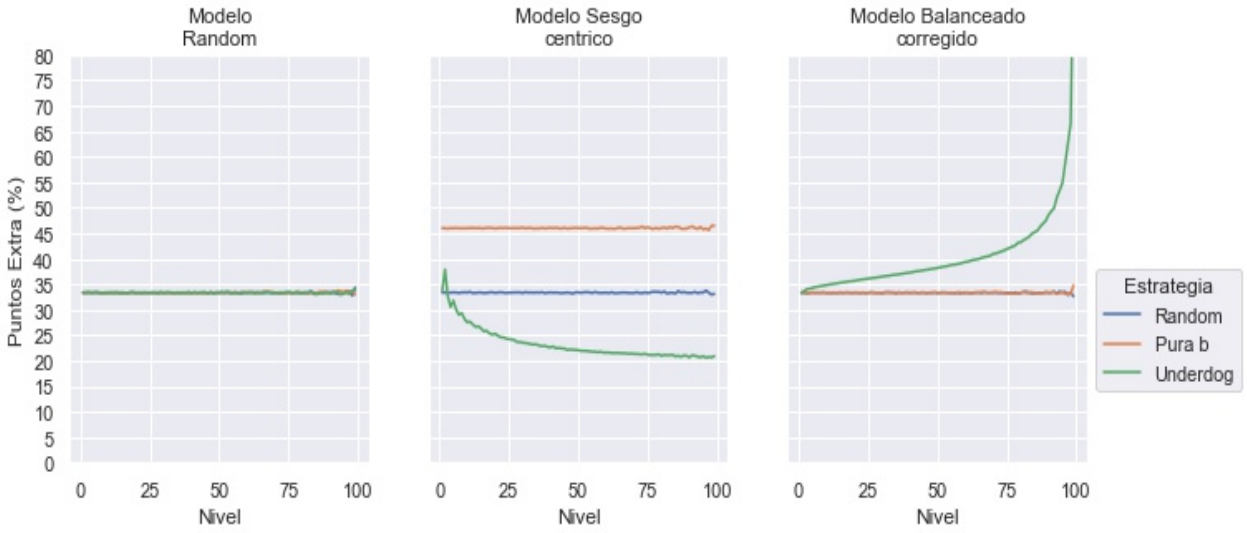
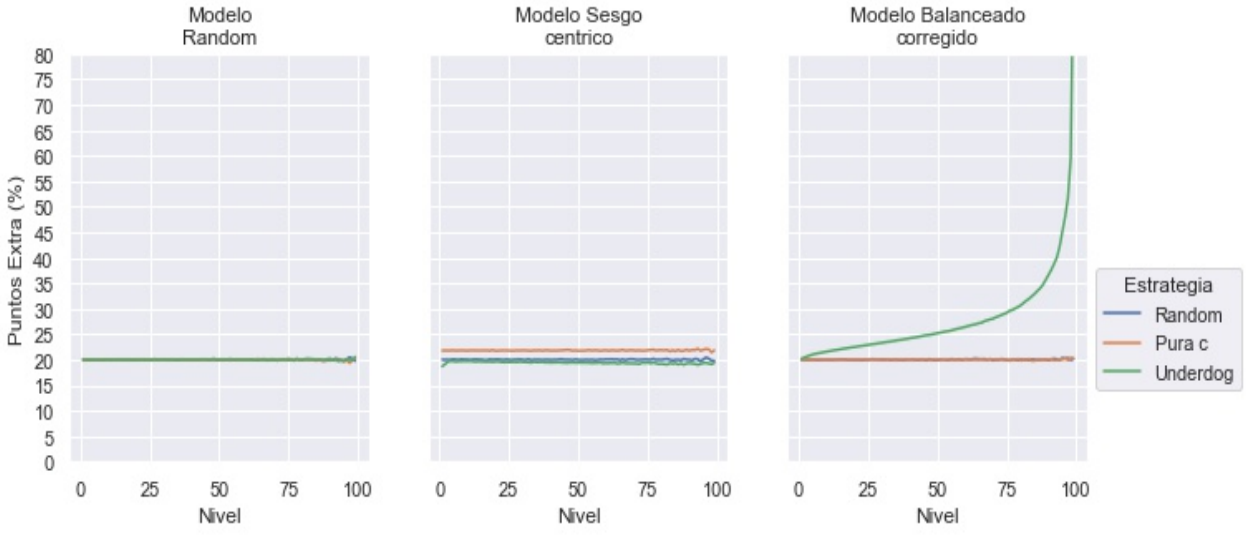


Figura 5.39: Estudio 1 en 5 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.



Las ganancias porcentuales para distintos números de ítems en 3 y 5 opciones se presentan en las Figuras 5.40 y 5.41, respectivamente. Se replican las observaciones hechas para pruebas de 4 opciones con pruebas de 3 y 5 opciones: 1) el número de ítems no tiene influencia en la ganancia porcentual de la estrategia Pura C o Pura B y 2) la ganancia porcentual de la estrategia Underdog aumenta drásticamente en pruebas cuyo número de ítems es bajo. Además, se observa otro efecto no visto en 4 opciones, y es que en 3 opciones, la desventaja porcentual de Underdog en Sesgo Céntrico es mayor cuando aumenta el número de ítems, apoyando la idea que hay efectos en 3 opciones que se pierden en 4 y 5 opciones y mostrando que la ganancias de estrategias en el modelo Sesgo Céntrico sí podrían depender del número de ítems usado.

Figura 5.40: Estudio 1 en 3 opciones: Ganancia porcentual para número de ítems en $\{3, 6, ..., 99\}$

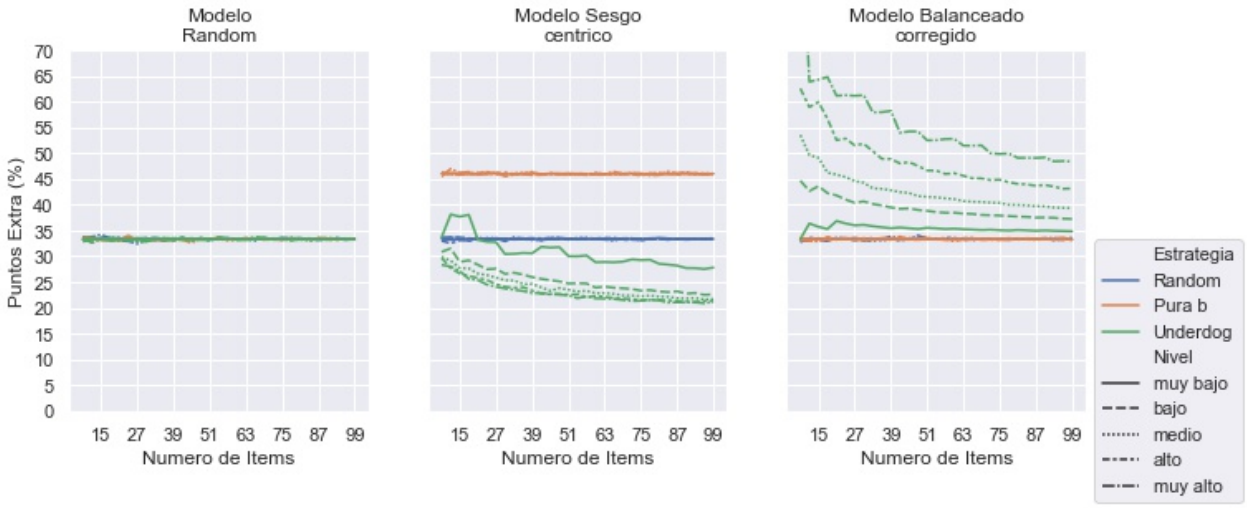
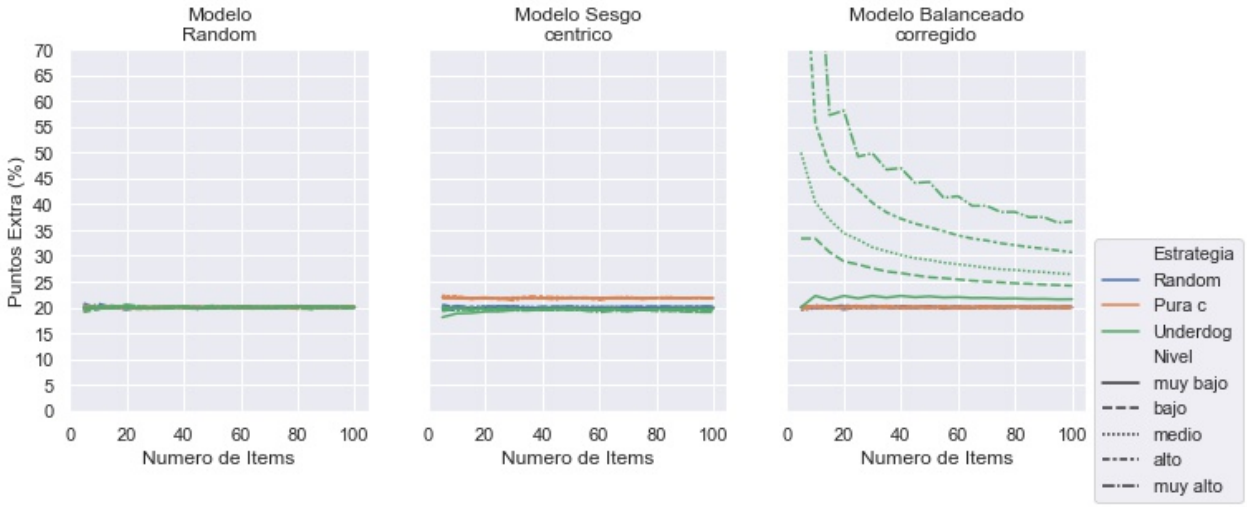


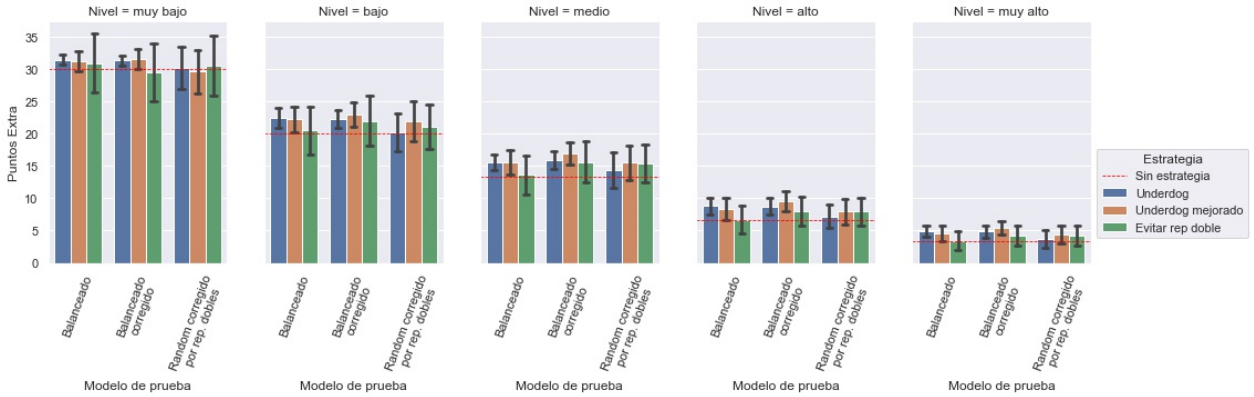
Figura 5.41: Estudio 1 en 5 opciones: Ganancia porcentual para número de ítems en $\{5, 10, ..., 100\}$



5.2.2. Replicación del Estudio 2

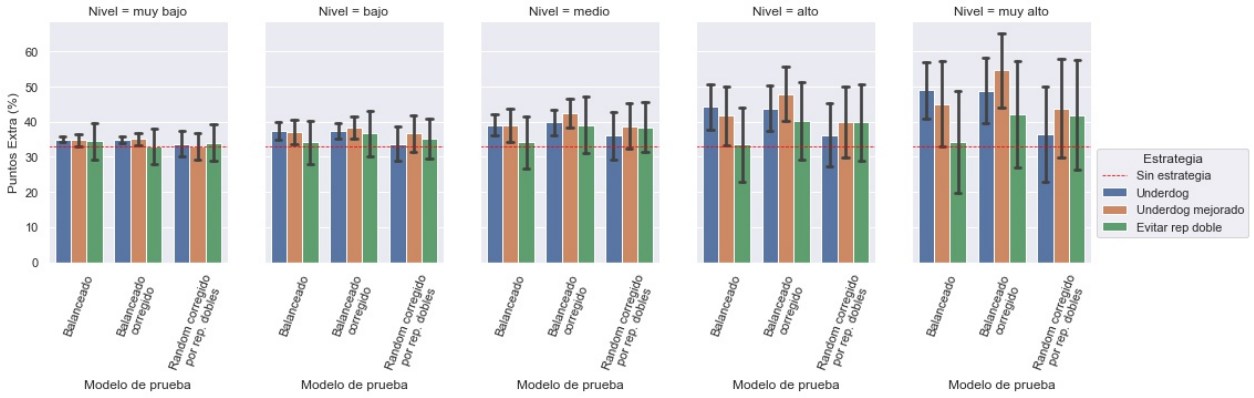
La ganancia y la ganancia porcentual obtenidas para pruebas de 100 ítems y examinados separados en 5 niveles, para los modelos Balanceado, Balanceado Corregido y Random Corregido por repeticiones dobles, y las estrategias Random, Underdog, Underdog Mejorado y Evitar Repeticiones Dobles son presentadas en las Figuras 5.42 y 5.43 para pruebas de 3 opciones y en las Figuras 5.44 y 5.45 para pruebas de 5 opciones.

Figura 5.42: Estudio 2 en 3 opciones: Ganancia para distintos modelos, estrategia y niveles.



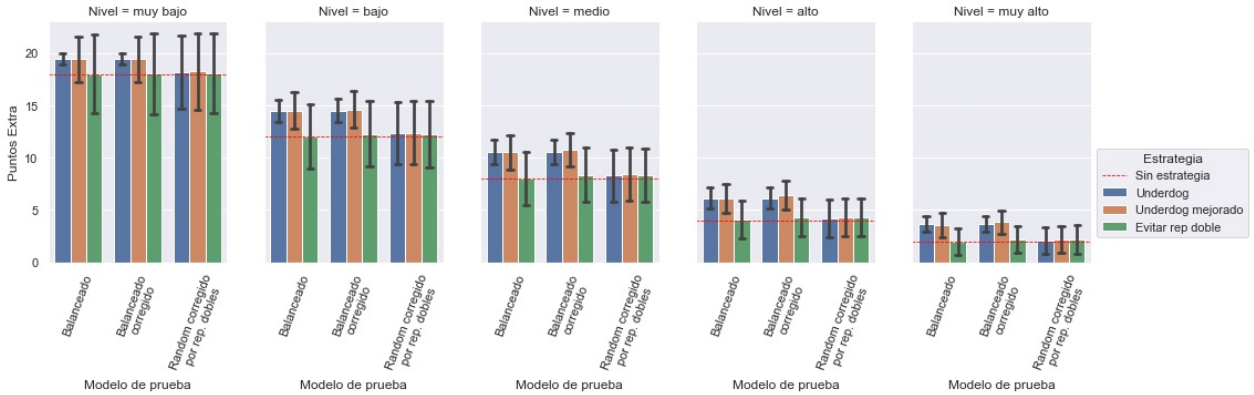
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.43: Estudio 2 en 3 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



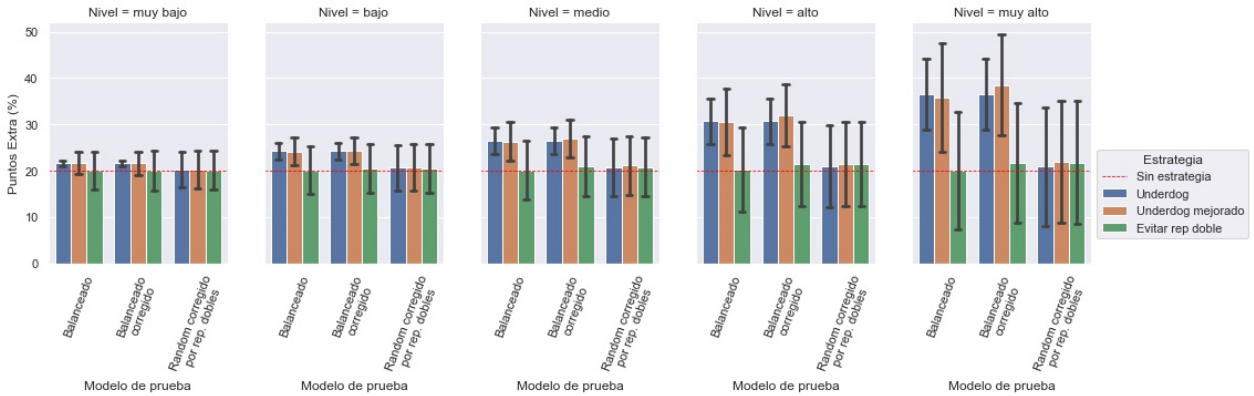
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.44: Estudio 2 en 5 opciones: Ganancia para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.45: Estudio 2 en 5 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Con respecto a la estrategia Underdog, al igual que en 4 opciones, obtuvo mayor ganancia que Random en el modelo Balanceado Exacto (3 opciones: Δ_{10}=1.4, d_{10}=0.4; Δ_{40}=2.4, d_{40}=0.9; Δ_{60}=2.4, d_{60}=1.1; Δ_{80}=2.0, d_{80}=1.2; Δ_{90}=1.5, d_{90}=1.2; p-values < 0.0001; 5 opciones: Δ_{10}=1.3, d_{10}=0.4; Δ_{40}=2.4, d_{40}=0.8; Δ_{60}=2.5, d_{60}=1.0; Δ_{80}=2.1, d_{80}=1.2; Δ_{90}=1.5, d_{90}=1.2; p-values < 0.0001) y obtuvo igual ganancia entre los modelos Balanceado Exacto y Balanceado Corregido (p-values > 0.1). Para el modelo Random Corregido Por Repeticiones Dobles, en 5 opciones, al igual que en 4 opciones, la ganancia de Underdog fue similar (tamaños de efecto despreciables) a Random (Δ_{10}=0.3; Δ_{40}=0.3; Δ_{60}=0.3; Δ_{80}=0.2; Δ_{90}=0.1; p-values < 0.0001; d_{10}, d_{40}, d_{60}, d_{80}, $d_{90} = 0.1$) pero en 3 opciones la ganancia de Underdog fue levemente mayor (tamaños de efecto desde despreciables hasta pequeños) o igual a Random, (Δ_{10}=0.6, d_{10}=0.1; Δ_{40}=0.8, d_{40}=0.2; Δ_{60}=0.8, d_{60}=0.3; Δ_{80}=0.4, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p-values < 0.0001) La pequeña ventaja de Underdog para 3 opciones en Random Corregido por repeticiones dobles puede ser explicada viendo que evitar repeticiones a nivel local provoca un leve balanceo de las claves a nivel global (que es mayor en 3 opciones), y este leve balanceo es aprovechado por la estrategia Underdog.

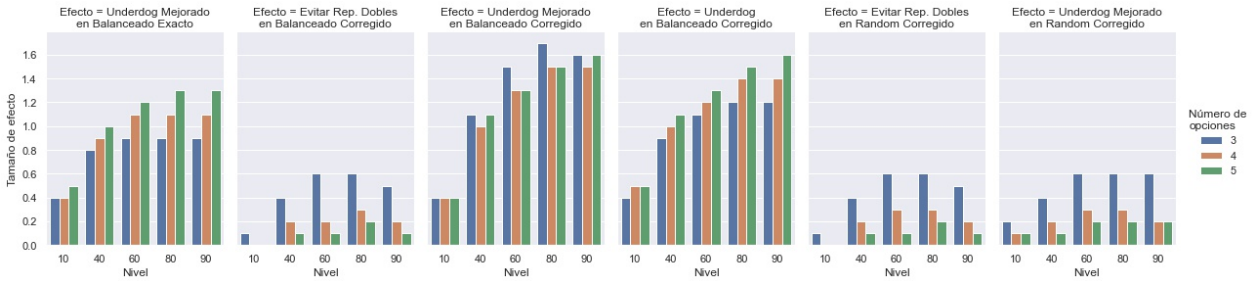
Con respecto a la comparación de Underdog con Underdog mejorado en 3 opciones, al igual que en pruebas de 4 opciones, la ganancia de Underdog Mejorado fue mayor para niveles mayores que 10 e igual en el nivel 10 que la ganancia de Underdog en el modelo Balanceado Corregido (Δ_{40}=0.7, d_{40}=0.4; Δ_{60}=1.1, d_{60}=0.7; Δ_{80}=1.1, d_{80}=0.8; Δ_{90}=0.7, d_{90}=0.6; p_{40},p_{60},p_{80},$p_{90} < 0.0001$, $p_{10} = 0.8$), y al contrario, en el modelo Balanceado Exacto, la ganancia de Underdog Mejorado fue levemente menor o igual que la ganancia de Underdog (Δ_{10}=0.0, d_{10}=0.0, p_{10}=1.0; Δ_{40}=0.1, d_{40}=0.0; Δ_{60}=0.1, d_{60}=0.1; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.3, d_{90}=0.3; p_{40},p_{60},p_{80},$p_{90} < 0.0001$). Para 5 opciones, al igual que en pruebas de 4 opciones, la ganancia de Underdog Mejorado fue mayor o igual que la ganancia de Underdog en el modelo Balanceado Corregido, pero este efecto fue pequeño en todos los niveles (Δ_{40}=0.1; Δ_{60}=0.2; Δ_{80}=0.2; Δ_{90}=0.2; d_{40}, d_{60}, d_{80}, $d_{90} < 0.2$; p_{40}, p_{60}, p_{80}, $p_{90} < 0.0001$, p_{10}=0.6), y para el modelo Balanceado Exacto, las ganancias de Underdog Mejorado y Underdog fueron iguales (p-values > 0.2). Estos resultados podrían sugerir que los efectos que se ven son los mismos, pero atenúan la magnitud del efecto cuando el número de opciones es más grande.

En el modelo Random Corregido por repeticiones dobles, la ganancia de la estrategia Evitar Repeticiones Dobles replicó los efectos en 4 opciones, fue levemente mayor (tamaños de efecto hasta mediano en el caso de 3 opciones, y hasta pequeños en el caso de 5 opciones) a la ganancia de la estrategia Random (3 opciones: Δ_{10}=0.4, d_{10}=0.1; Δ_{40}=1.5, d_{40}=0.4; Δ_{60}=1.9, d_{60}=0.6; Δ_{80}=1.4, d_{80}=0.6; Δ_{90}=0.8, d_{90}=0.5; p-values < 0.0001; 5 opciones:Δ_{40}=0.2, d_{40}=0.1; Δ_{60}=0.3, d_{60}=0.1; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.1; p_{40},p_{60},p_{80},$p_{90} < 0.0001$, p_{10}=0.3). Por otro lado, en el caso de la ganancia de Evitar Repeticiones dobles en Balanceado Corregido, en 5 opciones, al igual que en 4 opciones, fue similar (tamaños de efecto despreciables) a Random (Δ_{10}=0.1, d_{10}=0.0; Δ_{40}=0.3, d_{40}=0.1; Δ_{60}=0.3, d_{60}=0.1; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.1; p-values < 0.0001), pero en 3 opciones fue levemente mayor (tamaños de efecto hasta pequeños) a Random (Δ_{10}=0.4, d_{10}=0.1; Δ_{40}=1.6, d_{40}=0.4; Δ_{60}=1.8, d_{60}=0.6; Δ_{80}=1.4, d_{80}=0.6; Δ_{90}=0.8, d_{90}=0.5; p-values < 0.0001).

En la comparación de tamaños de efecto presentada en la Figura 5.46, se puede notar cómo el número de opciones influye diferentemente en la ganancia de la estrategia Underdog Mejorado (con respecto al control Random) según el modelo de construcción en estudio: Por un lado, en el modelo Balanceado Exacto los tamaños de efecto son más grandes cuando el número de opciones es más grande; por otro lado, lo contrario ocurre en el modelo Random Corregido Por Repeticiones Dobles; finalmente, cuando el modelo es la combinación de ambos (Balanceado Corregido), los tamaños de efecto no parecen ser más grandes ni mas chicos con respecto al número de opciones. Esto podría indicar que los efectos del uso de una estrategia (al menos en un modelo que evita repeticiones y balancea) aumenta su ventaja (con respecto al número de opciones) cuando saca provecho del balanceo, y disminuye su ventaja cuando saca provecho de la evitación de repeticiones, siendo la última comparación de los efectos ambigua porque sería una combinación de un crecimiento y un decrecimiento.

Revisando el caso de la estrategia Evitar Repeticiones Dobles, se puede ver en la Figura 5.46 que, en ambos modelos donde saca ventaja, los tamaños de efecto aumentan consistentemente cuando el número de opciones decrece. Por su parte, en el caso de la estrategia Underdog (que saca provecho del balanceo), para el modelo Balanceado Corregido (modelo que evita repeticiones) los efectos decrecen cuando el número de opciones decrece pero en modelos que no evitan repeticiones como Balanceo Exacto, no hay una tendencia clara. Podemos resumir los resultados de la comparación de los tamaños de efecto como sigue: En

Figura 5.46: Estudio 2: Comparación de los tamaños de efecto encontrados en distinto número de opciones.



el caso de los modelos que evitan repeticiones, cuando el número de opciones aumenta, los tamaños de efectos 1) crecen si la estrategia saca provecho del balanceo y 2) disminuyen si la estrategia saca provecho de la evitación de repeticiones.

Se observa en la ganancia porcentual que en los modelos Balanceado y Balanceado Corregido la ganancia porcentual de las estrategias Underdog y Underdog Mejorado, al igual que en 4 opciones, son crecientes con respecto al nivel del examinado en 3 y 5 opciones. La ganancia porcentual de Evitar Repeticiones Dobles en Balanceado Corregido y Random Corregido por repeticiones dobles, y la ganancia porcentual de la estrategia Underdog Mejorado en Random Corregido por repeticiones dobles y Evitar Repeticiones Dobles no se ven afectadas mayormente por el nivel de desempeño en 5 opciones (al igual que en 4 opciones), pero son crecientes con respecto al nivel del examinado en 3 opciones.

En la Figura 5.47 y 5.48, se presenta la ganancia porcentual en todos los niveles de desempeño posibles para pruebas de 3 y 5 opciones, respectivamente. Se confirman las observaciones descriptivas de la ganancia porcentual: En los modelos Balanceado y Balanceado Corregido, la ganancia porcentual de Underdog y Underdog Mejorado es exponencialmente creciente con respecto al nivel. Con Evitar Repeticiones en los modelos Balanceado Corregido y Random Corregido Por Repeticiones Dobles y Underdog Mejorado en el modelo Random Corregido Por Repeticiones Dobles la ganancia porcentual es linealmente creciente con una pendiente considerable en el caso de 3 opciones, y constante en el caso de 5 (y 4) opciones.

Figura 5.47: Estudio 2 en 3 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.

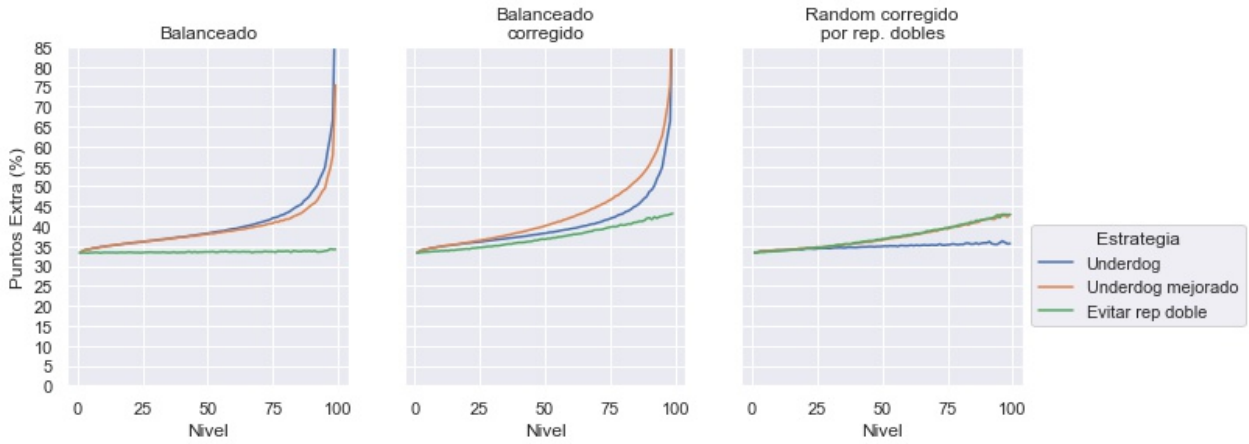
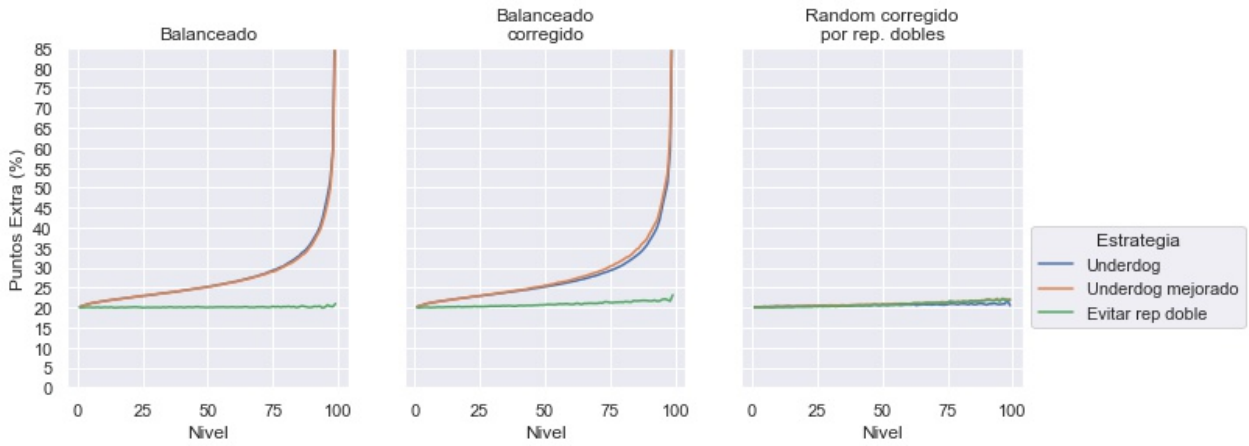


Figura 5.48: Estudio 2 en 5 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.



La ganancia porcentual para distinto número de ítems se presenta en la Figura 5.49 y 5.50. Se observa que el número de ítems no tiene influencia en la ganancia porcentual de las estrategias Underdog Mejorado y Evitar Repeticiones Dobles en los modelos Balanceado Corregido y Random Corregido por repeticiones dobles. Por otra parte, la ganancia porcentual de las estrategias Underdog y Underdog mejorado, en los modelos Balanceado y Balanceado Corregido crece drásticamente en pruebas cuyo número de ítems es bajo. Se puede notar también que en 3 opciones la diferencia entre la ganancia porcentual de Underdog Mejorado y Underdog en el modelo Balanceado Corregido crece mientras mayor es el número de ítems, pero que esta diferencia en 5 opciones es menos notoria (aunque también crece con respecto al número de ítems).

Figura 5.49: Estudio 2 en 3 opciones: Ganancia porcentual para número de ítems en $\{3, 6, ..., 99\}$

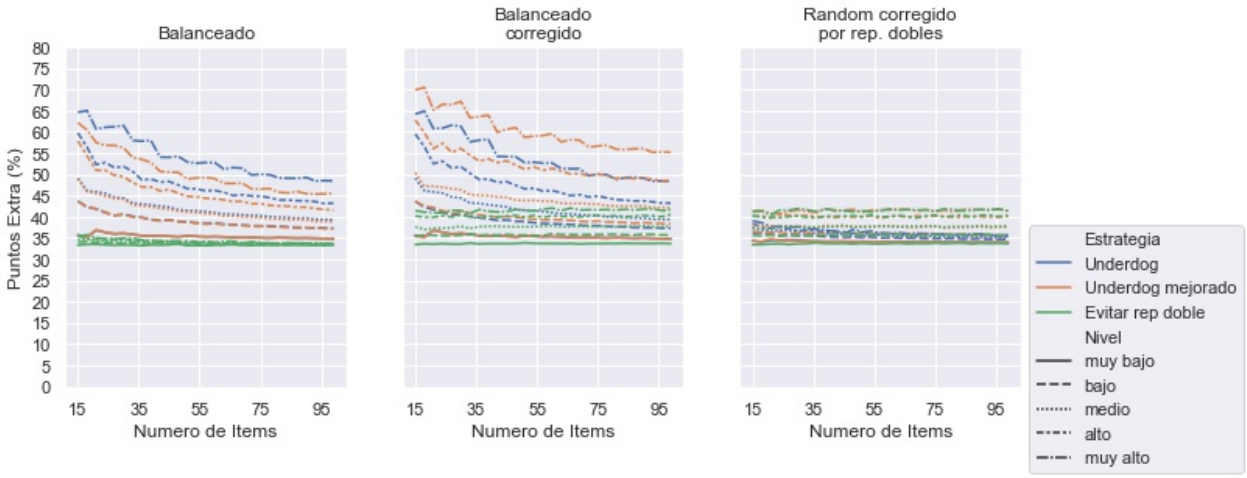
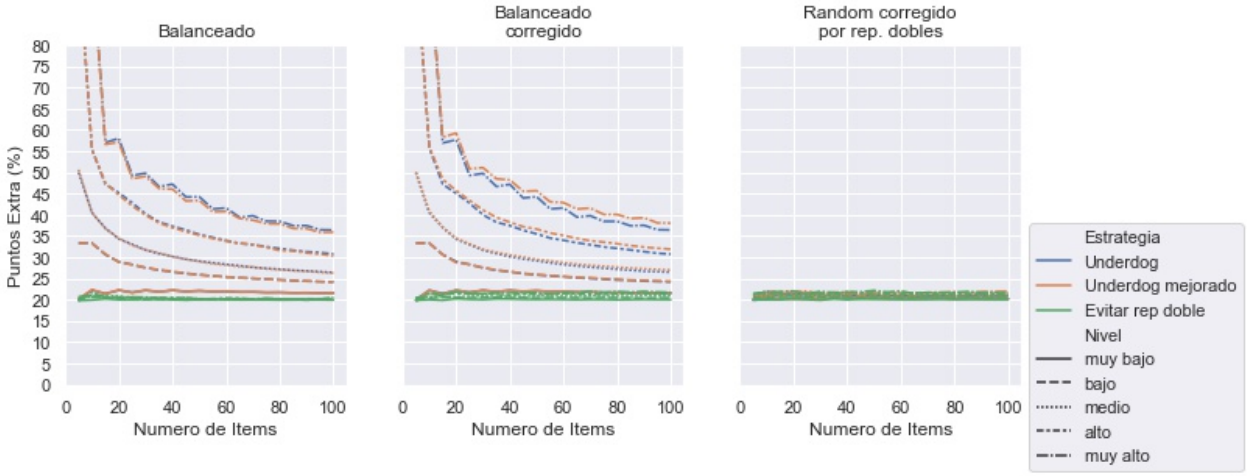


Figura 5.50: Estudio 2 en 5 opciones: Ganancia porcentual para número de ítems en $\{5, 10, ..., 100\}$



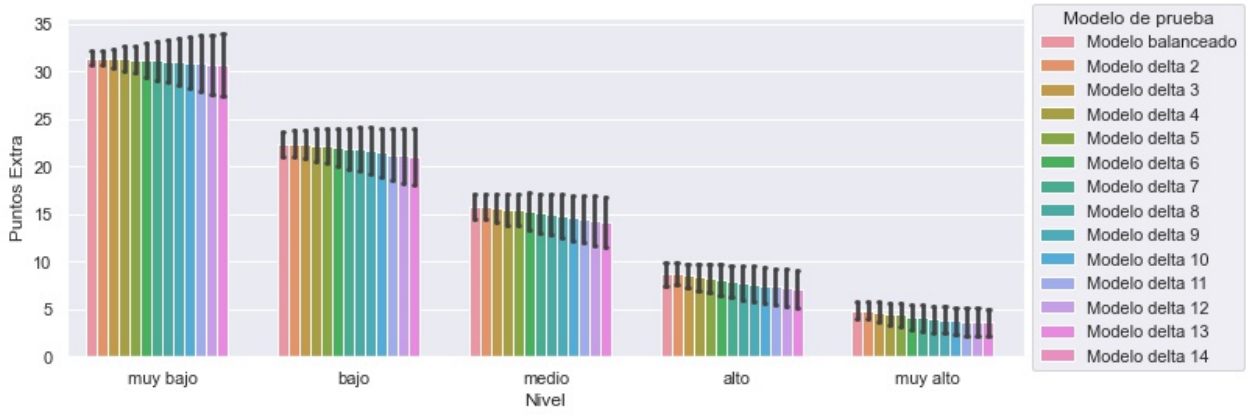
La ganancia y la ganancia porcentual obtenidas en pruebas de 100 ítems y estudiantes separados en 5 niveles, para la estrategia Underdog en los modelos Balanceado Exacto, Balanceado Aproximativo ($\delta \in \{2, ..., 14\}$) y Random son presentadas en las Figuras 5.51 y 5.53 para 3 opciones, y en las Figuras 5.52 y 5.54 para 5 opciones. Se replicaron todos los efectos del Estudio 2 observados en pruebas de 4 opciones para pruebas de 3 y 5 opciones también (incluyendo además de los efectos hasta $\delta = 8$, efectos hasta para un $\delta = 14$):

1. En todos los modelos (Balanceado Exacto y Balanceado Aproximativo) excepto en Random, la ganancia de Underdog fue significativamente mayor que la ganancia de la estrategia Random (3 opciones: Balanceado Exacto $\rightarrow \Delta_{10}$=1.4, d_{10}=0.4, p_{10}=0.0; Δ_{40}=2.4, d_{40}=0.9, p_{40}=0.0; Δ_{60}=2.4, d_{60}=1.1, p_{60}=0.0; Δ_{80}=2.0, d_{80}=1.2, p_{80}=0.0; Δ_{90}=1.5, d_{90}=1.2, p_{90}=0.0; Balanceado Aproximativo delta 14 $\rightarrow \Delta_{10}$=0.6, d_{10}=0.1, p_{10}=0.0;

Δ_{40}=0.7, d_{40}=0.2, p_{40}=0.0; Δ_{60}=0.7, d_{60}=0.2, p_{60}=0.0; Δ_{80}=0.4, d_{80}=0.2, p_{80}=0.0; Δ_{90}=0.2, d_{90}=0.1, p_{90}=0.0; 5 opciones: Balanceado Exacto → Δ_{10}=1.5, d_{10}=0.5, p_{10}=0.0; Δ_{40}=2.5, d_{40}=1.1, p_{40}=0.0; Δ_{60}=2.6, d_{60}=1.3, p_{60}=0.0; Δ_{80}=2.2, d_{80}=1.5, p_{80}=0.0; Δ_{90}=1.7, d_{90}=1.6, p_{90}=0.0; Balanceado Aproximativo delta 14 → Δ_{10}=0.2, d_{10}=0.1, p_{10}=0.0; Δ_{40}=0.4, d_{40}=0.1, p_{40}=0.0; Δ_{60}=0.3, d_{60}=0.1, p_{60}=0.0; Δ_{80}=0.2, d_{80}=0.1, p_{80}=0.0; Δ_{90}=0.1, d_{90}=0.1, p_{90}=0.0).

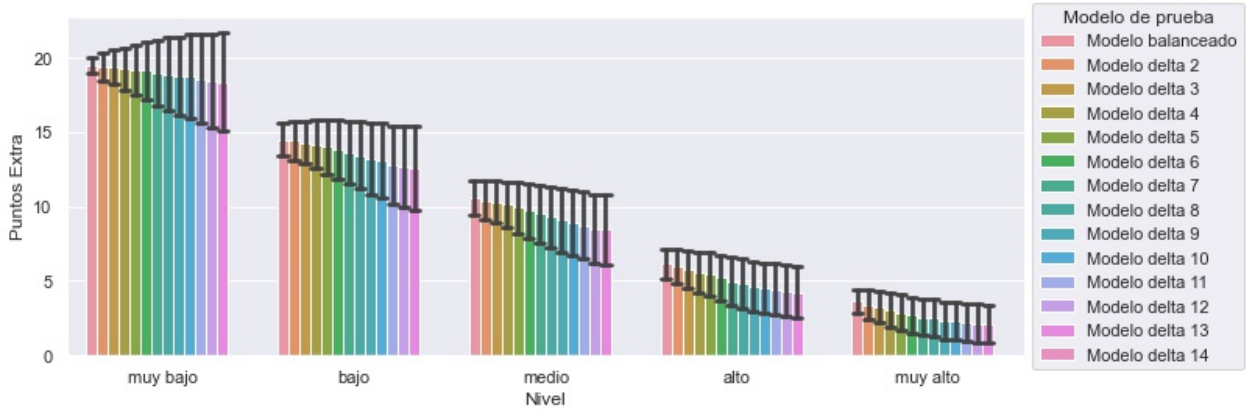
2. La ganancia de la estrategia Underdog en el modelo Balanceado Exacto fue significativamente mayor o similar a la ganancia en los modelos Balanceado Aproximativo. Además, la ganancia de Underdog fue decreciente con respecto a δ en los modelos balanceo Aproximativo. Los indicadores son omitidos porque son muchos y todos van en la dirección de lo antes expuesto.

Figura 5.51: Estudio 2 en 3 opciones: Ganancia de Underdog para los modelos Balanceado Aproximativo.



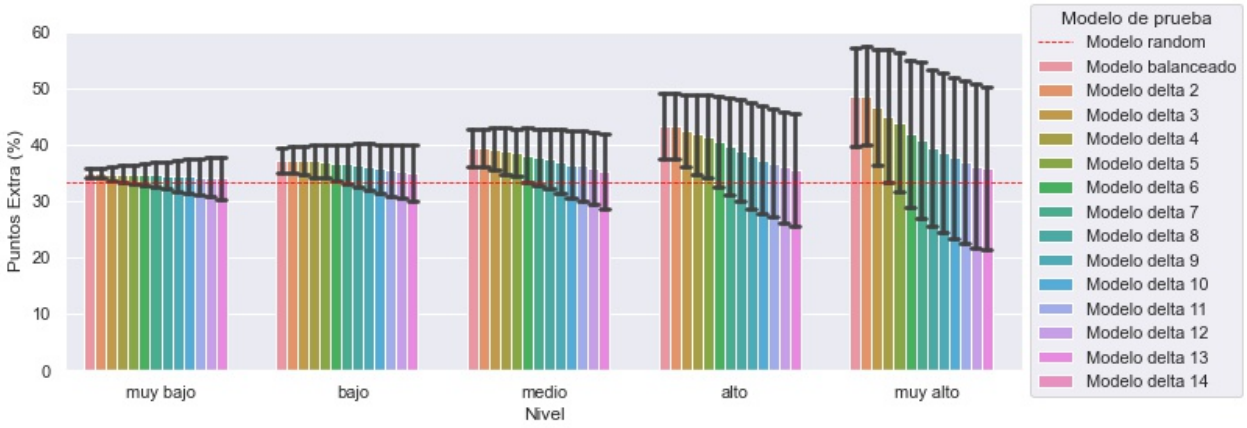
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.52: Estudio 2 en 5 opciones: Ganancia de Underdog para los modelos Balanceado Aproximativo.



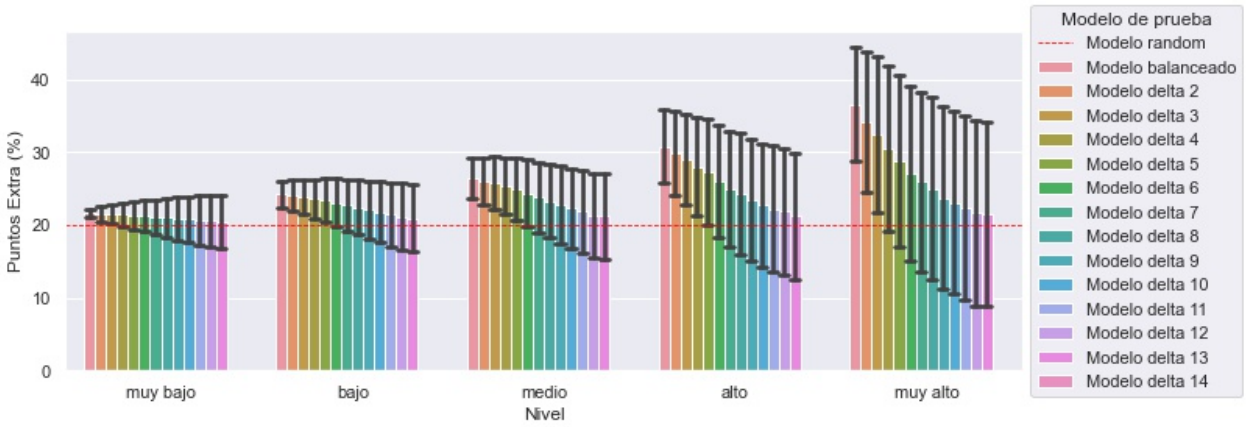
Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.53: Estudio 2 en 3 opciones: Ganancia porcentual de Underdog para los modelos Balanceado Aproximativo.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.54: Estudio 2 en 5 opciones: Ganancia porcentual de Underdog para los modelos Balanceado Aproximativo.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

La ganancia porcentual en los 5 niveles se presenta en las Figuras 5.53 y 5.54. Se observa que la ganancia porcentual de la estrategia Underdog en todos los modelos excepto el modelo Random es creciente con respecto al nivel, pero a medida que crece δ en los modelos Balanceado Aproximativo este crecimiento se hace cada vez menos notable.

Figura 5.55: Estudio 2 en 3 opciones: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para niveles en $\{1, ..., 99\}$.

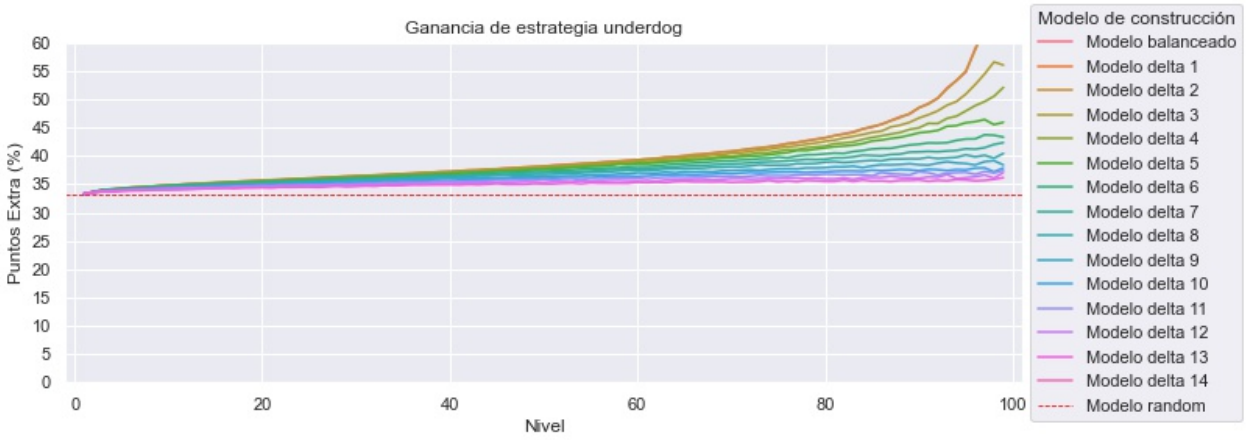
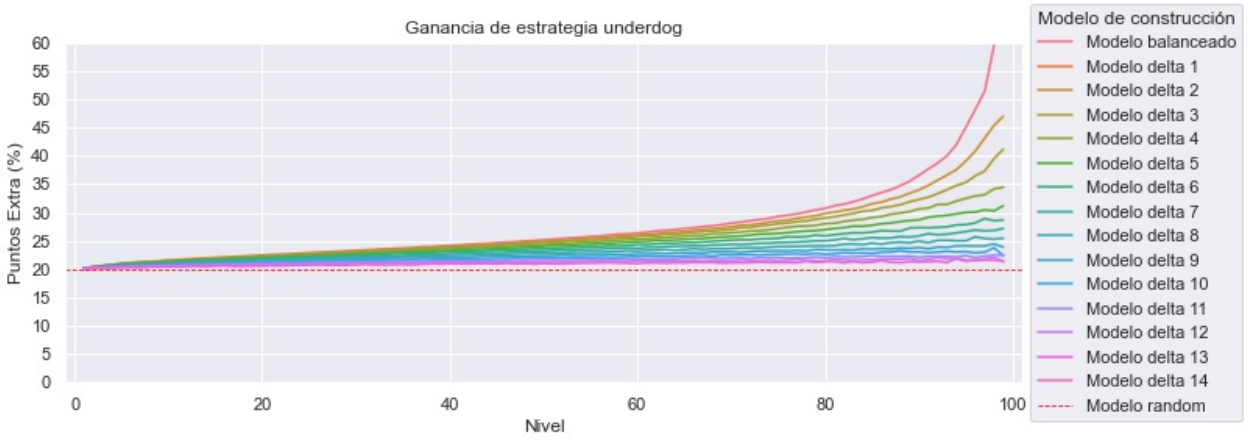


Figura 5.56: Estudio 2 en 5 opciones: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para niveles en $\{1, ..., 99\}$.



En la Figura 5.55 y 5.56 se presenta la ganancia porcentual en todos los niveles de desempeño posibles para 3 y 5 opciones, respectivamente. Confirmando la observación de la ganancia porcentual en 5 niveles, se puede ver que la ganancia porcentual de Underdog es creciente con respecto al nivel en todos los modelos excepto el modelo Random, además, el crecimiento pasa de ser exponencial a lineal a medida que el delta crece.

La ganancia porcentual para distinto número de ítems se presenta en la Figura 5.57 y 5.58. Se observa la ganancia porcentual de la estrategia Underdog aumenta drásticamente en pruebas cuyo número de ítems es bajo para todos los modelos excepto el Random. Además, cuando el número de ítems es bajo, la diferencia entre los distintos modelos decrece, lo que indica que el posible efecto regulador de la ganancia que tiene δ (es decir que mientras mayor sea delta menor es la ventaja de Underdog) parece desaparecer cuando el número de ítems es bajo. Tomando todos los resultados de los modelos Balanceado Aproximativo, se replicaron todos los efectos vistos en 4 opciones, y además, se encontraron efectos para deltas hasta 14.

Figura 5.57: Estudio 2 en 3 opciones: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para número de ítems en $\{3, 6, ..., 99\}$

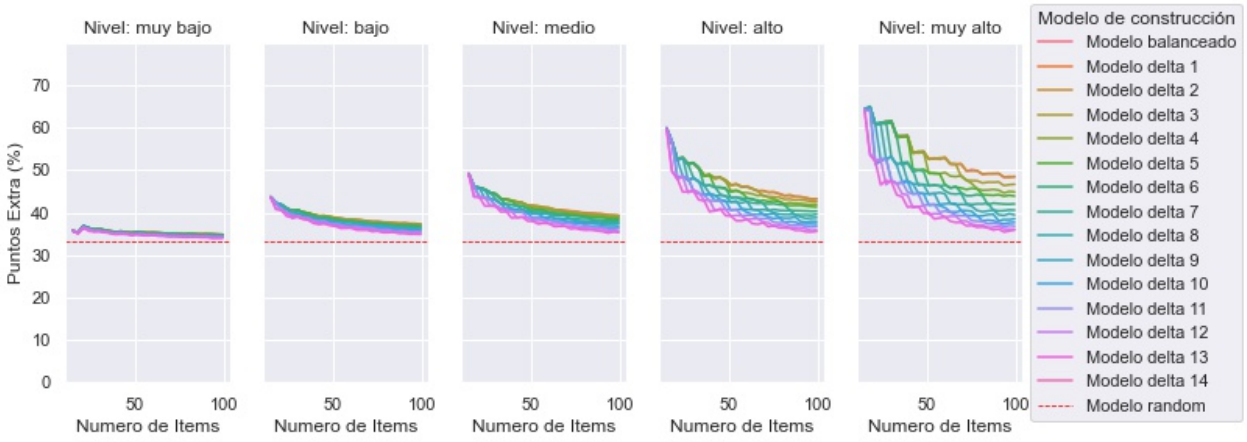
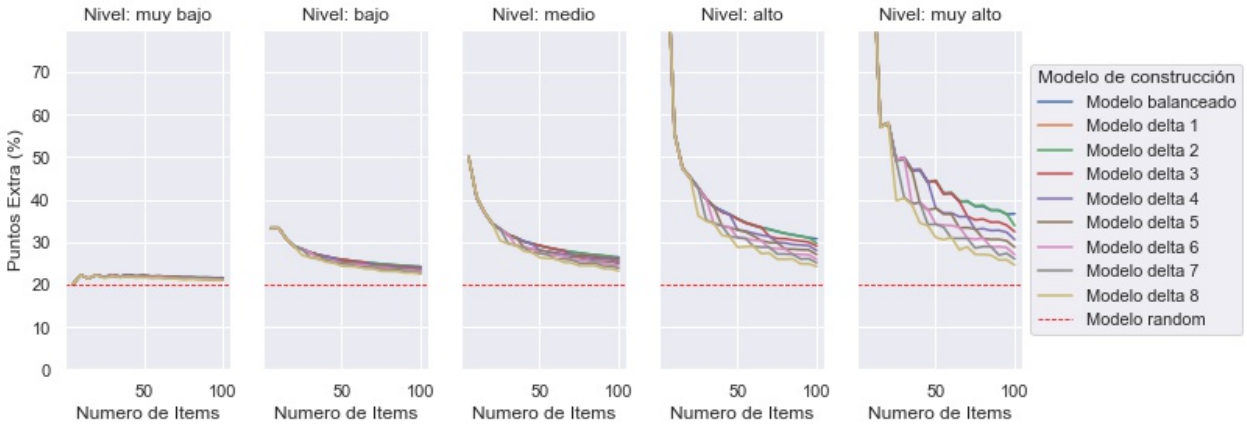


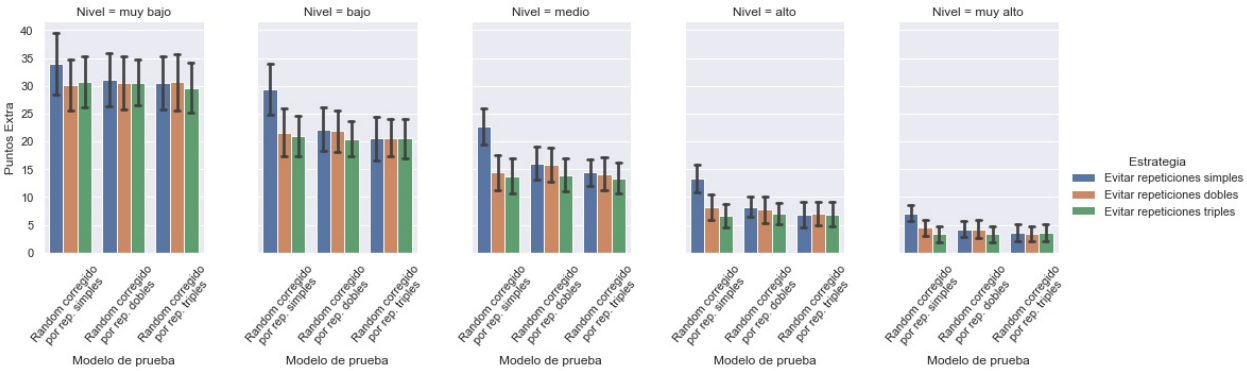
Figura 5.58: Estudio 2 en 5 opciones: Ganancias porcentuales obtenidas por Underdog en los modelos Balanceado Aproximativo para número de ítems en $\{5, 10, ..., 100\}$



5.2.3. Replicación del Estudio 3

La ganancia y la ganancia porcentual obtenidas en pruebas de 100 ítems y estudiantes separados en 5 niveles, para los modelos Random, Random Corregido por repeticiones simples, Random Corregido por repeticiones dobles, Random Corregido por repeticiones triples, y para las estrategias Random, Evitar Repeticiones Simples, Evitar Repeticiones Dobles, y Evitar Repeticiones Triples, son presentadas en las Figuras 5.59 y 5.60 para 3 opciones y en las Figuras 5.61 y 5.62 para 5 opciones.

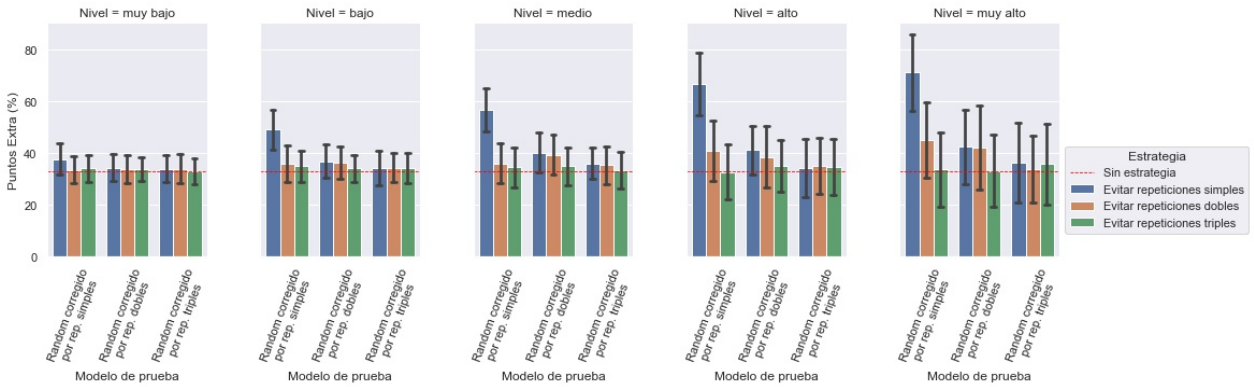
Figura 5.59: Estudio 3 en 3 opciones: Ganancia para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

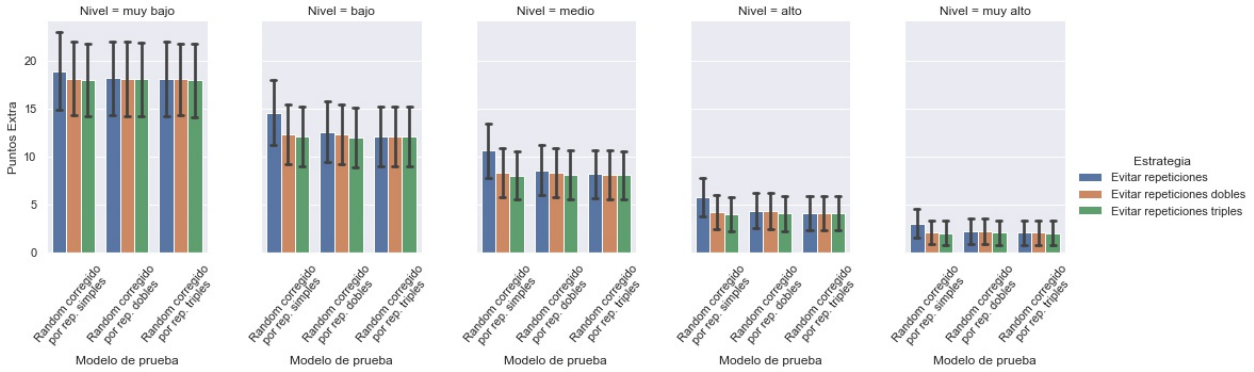
Para el modelo Random Corregido Por Repeticiones Simples, la ganancia de la estrategia Evitar Repeticiones Simples fue mayor a la de la estrategia Random para 3 y 5 opciones (3 opciones: Δ_{10}=3.2, d_{10}=0.6; Δ_{40}=9.5, d_{40}=2.3; Δ_{60}=9.7, d_{60}=2.9; Δ_{80}=6.6, d_{80}=3.0; Δ_{90}=3.7, d_{90}=2.5; p-values < 0.0001; 5 opciones: Δ_{10}=0.9, d_{10}=0.2; Δ_{40}=2.5, d_{40}=0.8; Δ_{60}=2.6, d_{60}=1.0; Δ_{80}=1.8, d_{80}=0.9; Δ_{90}=1.0, d_{90}=0.7; p-values < 0.0001), la ganancia de la estrategia Evitar Repeticiones Dobles fue similar a la de la estrategia Random en 5 opciones (Δ_{10}=0.2;

Figura 5.60: Estudio 3 en 3 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



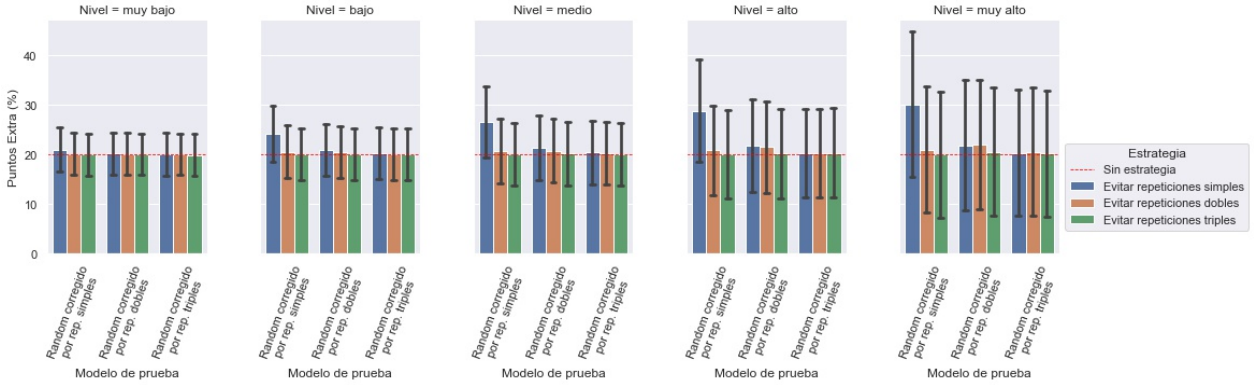
Nota Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.61: Estudio 3 en 5 opciones: Ganancia para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Figura 5.62: Estudio 3 en 5 opciones: Ganancia porcentual para distintos modelos, estrategia y niveles.



Nota: Las barras de error representan la desviación estándar con respecto al promedio.

Δ_{40}=0.3; Δ_{60}=0.3; Δ_{80}=0.2; Δ_{90}=0.1; p-values < 0.0001; d_{10}, d_{40}, d_{60}, d_{80}, d_{90} < 0.1) pero mayor (tamaños de efecto desde despreciables hasta medianos) para 3 opciones (Δ_{10}=0.5, d_{10}=0.1; Δ_{40}=1.4, d_{40}=0.4; Δ_{60}=1.6, d_{60}=0.5; Δ_{80}=1.2, d_{80}=0.5; Δ_{90}=0.7, d_{90}=0.5; p-values < 0.0001) y la ganancia de la estrategia Evitar Repeticiones Triples fue similar a la de la estrategia Random (3 opciones: Δ_{10}=0.1; Δ_{40}=0.5; Δ_{60}=0.4; Δ_{80}=0.2; Δ_{90}=0.0; p-values < 0.0001; d_{10}, d_{40}, d_{60}, d_{80}, d_{90} < 0.1; 5 opciones: p-values > 0.2). Estos resultados replican los mismos efectos vistos en 4 opciones, pero con mayor magnitud en el caso de 3 opciones y menor magnitud en 5 opciones.

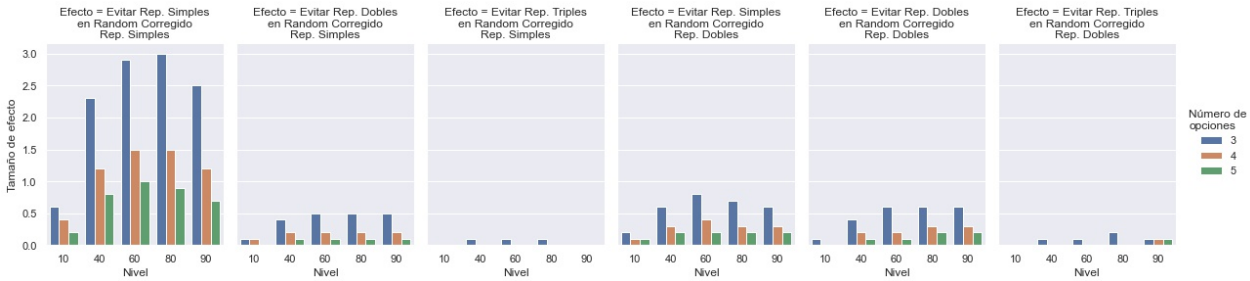
Para el modelo Random Corregido Por Repeticiones Dobles, al igual que en 4 opciones, las ganancias de Evitar Repeticiones Simples fueron mayores (tamaños de efecto desde despreciable a pequeños en 3 y 5 opciones) a Random (3 opciones: Δ_{10}=0.9, d_{10}=0.2; Δ_{40}=2.3, d_{40}=0.6; Δ_{60}=2.3, d_{60}=0.8; Δ_{80}=1.6, d_{80}=0.7; Δ_{90}=0.9, d_{90}=0.6; p-values < 0.0001 ; 5 opciones: Δ_{10}=0.2, d_{10}=0.1; Δ_{40}=0.5, d_{40}=0.2; Δ_{60}=0.6, d_{60}=0.2; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p-values < 0.0001), las ganancias de Evitar Repeticiones Dobles fueron mayores (tamaños de efecto desde despreciable a pequeños en 3 y 5 opciones) a Random (3 opciones: Δ_{10}=0.4, d_{10}=0.1; Δ_{40}=1.5, d_{40}=0.4; Δ_{60}=1.7, d_{60}=0.6; Δ_{80}=1.4, d_{80}=0.6; Δ_{90}=0.9, d_{90}=0.6; p-values < 0.0001 ; 5 opciones: Δ_{10}=0.1, d_{10}=0.0; Δ_{40}=0.3, d_{40}=0.1; Δ_{60}=0.3, d_{60}=0.1; Δ_{80}=0.3, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p-values < 0.0001) y las ganancias de Evitar Repeticiones Triples fueron similares (tamaños de efecto despreciables) a Random Δ_{10}=0.1; Δ_{40}=0.3; Δ_{60}=0.3; Δ_{80}=0.4; Δ_{90}=0.2; p-values < 0.0001; d_{10}, d_{40}, d_{60}, d_{80}, d_{90} < 0.2 ; 5 opciones: Δ_{90}=0.1, d_{90}=0.1, p_{90}=0.0; p_{10}, p_{40}, p_{60}, p_{80} > 0.1).

Para el modelo Random Corregido Por Repeticiones Triples, las ganancias en 5 opciones, al igual que en 4 opciones, fueron similares a Random (Evitar Repeticiones Simples: Δ_{10}=0.2; Δ_{40}=0.2; Δ_{60}=0.1; Δ_{80}=0.1; Δ_{90}=0.0; p-values < 0.0001; d_{10}, d_{40}, d_{60}, d_{80}, d_{90} = 0.0; Evitar Repeticiones Dobles: d_{10}, d_{40}, d_{60}, d_{80}, d_{90}=0.0; Evitar Repeticiones Triples: Δ_{80}=0.1, d_{80}=0.0, p_{80}=0.0; p_{10}, p_{40}, p_{60}, p_{90} > 0.1). Por otra parte, en 3 opciones todas las estrategias logran sacar al menos una pequeña ventaja con tamaño de efecto 0.2 en cada caso (Evitar Repeticiones Simples: Δ_{10}=0.4, d_{10}=0.1; Δ_{40}=1.0, d_{40}=0.3; Δ_{60}=1.0, d_{60}=0.3; Δ_{80}=0.6, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.2; p-values < 0.0001; Evitar Repeticiones Dobles: Δ_{10}=0.1, d_{10}=0.0, p_{10}=0.4; Δ_{40}=0.6, d_{40}=0.2; Δ_{60}=0.6, d_{60}=0.2; Δ_{80}=0.6, d_{80}=0.3; Δ_{90}=0.3, d_{90}=0.2; p_{40},p_{60},p_{80},p_{90} < 0.0001; Evitar Repeticiones Triples: Δ_{10}=0.1, d_{10}=0.0, p_{10}=0.3; Δ_{40}=0.3, d_{40}=0.1; Δ_{60}=0.5, d_{60}=0.2; Δ_{80}=0.4, d_{80}=0.2; Δ_{90}=0.2, d_{90}=0.2; p_{40},p_{60},p_{80},p_{90} < 0.0001). Nuevamente, estos resultados replican los mismos efectos vistos en 4 opciones, pero con mayor magnitud en el caso de 3 opciones (haciendo aparecer efectos en el modelo para las 3 estrategias) y menor magnitud en 5 opciones.

Con respecto a los tamaños de efectos mostrados en la Figura 5.63, para todas las ganancias encontradas los efectos disminuyen cuando el número de opciones aumenta. La dirección consistente de estos efectos (aunque similar en algunos casos, siempre con la misma dirección) apoya la idea de que cuando el modelo evita repeticiones de algún tipo, y la estrategia saca ventaja de la evitación de repeticiones (como en todas las estrategias del Estudio 3), entonces el tamaño de efecto para 3 opciones es mayor que para 4 opciones, y mayor en 4 opciones que en 5 opciones.

Se observa en las Figuras 5.60 y 5.62 que la ganancia porcentual de la estrategia Evitar Repeticiones Simples es creciente con respecto al nivel. No obstante, para todos los demás

Figura 5.63: Estudio 3: Comparación de los tamaños de efecto encontrados en distinto número de opciones.



casos la ganancia es similar en todos los niveles. En la Figura 5.64 y 5.65 se presenta la ganancia porcentual en todos los niveles de desempeño posibles para pruebas de 3 y 5 opciones, respectivamente.

Figura 5.64: Estudio 3 en 3 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.

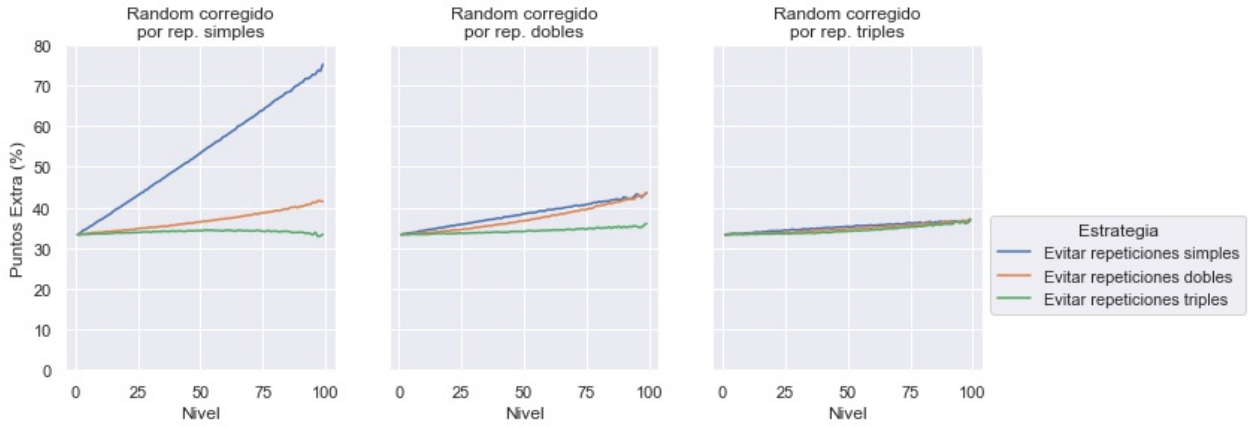
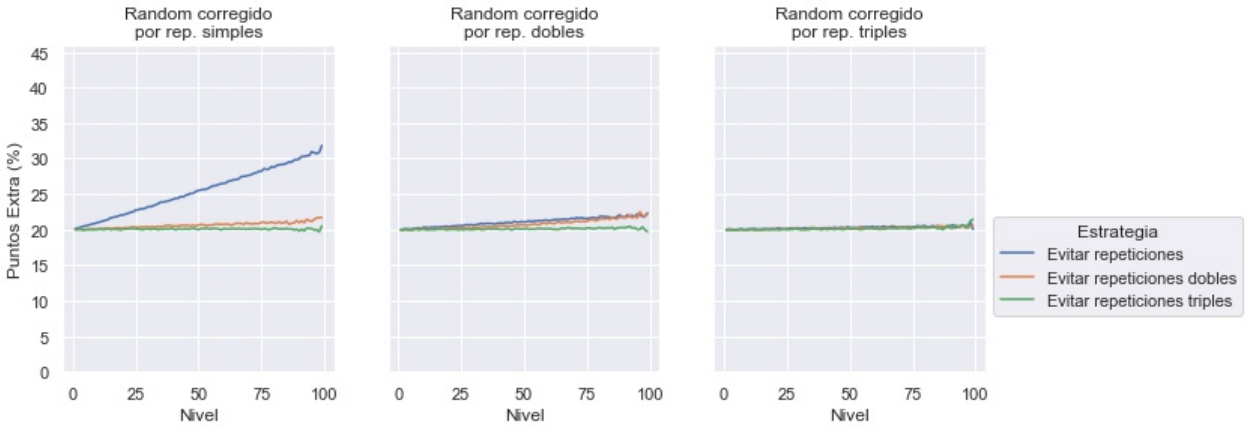


Figura 5.65: Estudio 3 en 5 opciones: Ganancia porcentual para niveles en $\{1, ..., 99\}$.



Confirmando la observación de la ganancia porcentual en 5 niveles, se puede ver que la ganancia porcentual de Evitar Repeticiones Simples crece linealmente con una pendiente alta, mientras que las otras ganancias también crecen linealmente pero con una pendiente muy baja.

La ganancia porcentual para distinto número de ítems se presenta en la Figura 5.66 y 5.67. Se observa que en todos los casos, la ganancia porcentual no es afectada por el número de ítems.

Figura 5.66: Estudio 3 en 3 opciones: Ganancia porcentual para número de ítems en $\{3, 6, ..., 99\}$

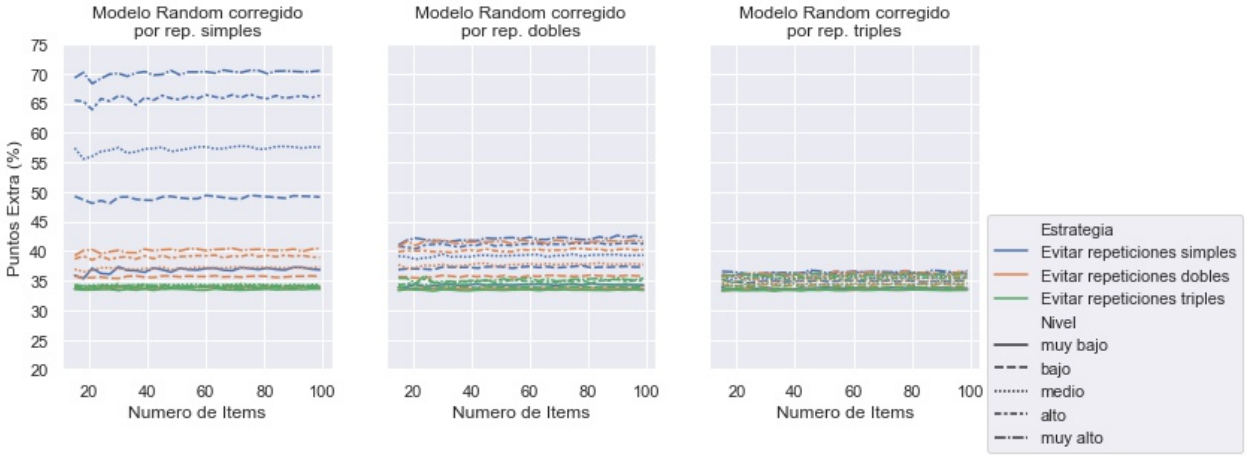
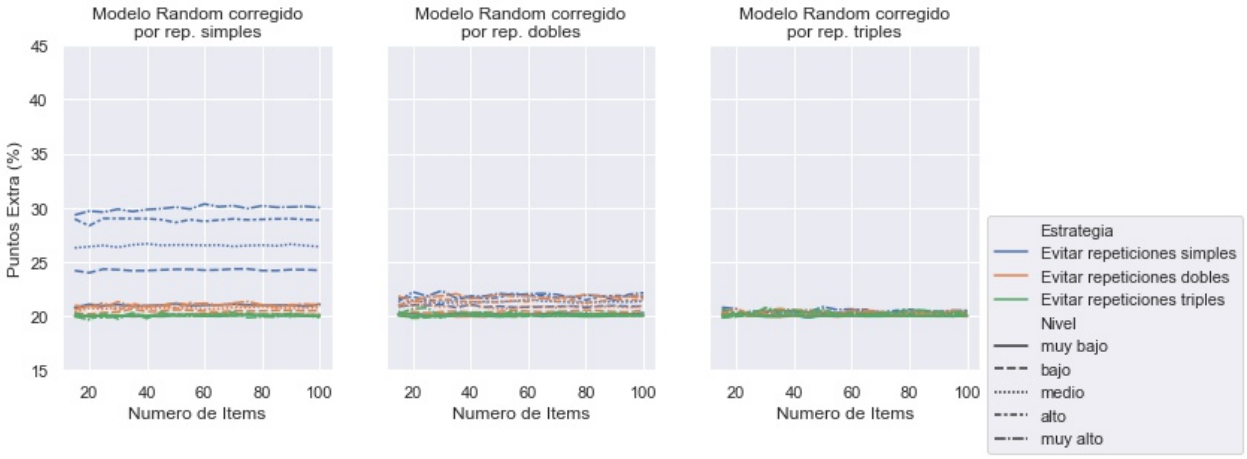


Figura 5.67: Estudio 3 en 5 opciones: Ganancia porcentual para número de ítems en $\{5, 10, ..., 100\}$



5.3. Discusión

Los resultados del Estudio 4 han podido replicar en gran medida los resultados de los Estudios 1, 2 y 3 obtenidos con pruebas de 4 opciones, para pruebas de 3 y 5 opciones.

Respecto al Estudio 1, se replicaron las ganancias de las estrategias Pura C y Underdog en el modelo Sesgo Céntrico y Balanceado Corregido, respectivamente. También se replicó la desventaja de puntos sufridas por el uso de Underdog en el modelo Sesgo Céntrico.

La ventaja de Pura C y la desventaja de Underdog en el modelo Sesgo Céntrico tuvieron un tamaño de efecto más fuerte en 3 que en 4 opciones, y mayor en 4 opciones que en 5 opciones. Para el caso de Pura C (o B) esto se podría deber a que el sesgo céntrico presente en las distribuciones de opciones correctas reportadas en la bibliografía es más fuerte en 3 opciones que en 4, y en 4 opciones que en 5 ($\frac{45{,}9\,\%}{33{,}3\,\%} = 1{,}37$ vs. $\frac{28{,}8\,\%}{25\,\%} = 1{,}15$ vs. $\frac{21{,}7\,\%}{20\,\%} = 1{,}08$). En el caso de la desventaja de Underdog, los tamaños de efecto entre número de opciones podrían ser explicados por dos observaciones: 1) la desventaja de Underdog se debe a que las claves encontradas por conocimiento tienen una frecuencia similar a la frecuencia dada por el sesgo céntrico (porque en el modelo Sesgo Céntrico las frecuencias de las claves en cada ítem son independientes e identicamente distribuidas), y con ello Underdog elige, en general, la clave que tenga menor frecuencia en el sesgo céntrico, la que es menor que el nivel Random para los 3 números de opciones, y por ello constituye una pérdida de puntos, 2) las frecuencias mínimas (i.e., las frecuencias de la mayoría de las claves que elige Underdog) relativas a random eran más bajas en 3 opciones que 4 opciones y que 5 opciones ($\frac{20{,}2\,\%}{33{,}3\,\%} = 0{,}60$ vs. $\frac{19{,}4\,\%}{25\,\%} = 0{,}77$ vs. $\frac{17{,}9\,\%}{20\,\%} = 0{,}89$). Estos dos puntos explican que la desventaja de Underdog sea mayor en 3 que en 4 opciones y mayor en 4 que en 5 opciones. Estos resultados muestran que no preocuparse por el posicionamiento de las opciones correctas podría inyectar más

ruido en la medición cuando los constructores usan ítems con pocas opciones.

La ventaja de Underdog en el modelo Balanceo Corregido también fue afectada por el número de opciones, pero en el sentido contrario, es decir, el efecto fue más fuerte en 5 opciones que en 4 opciones, y más fuerte en 4 opciones que en 3 opciones. Estos resultados muestran que balancear pone en riesgo la validez de los resultados de las pruebas para los 3 números de opciones, y afecta más a las pruebas con mayor número de opciones.

Con respecto a la influencia del nivel de desempeño, al igual que en caso de 4 opciones, el modelo Balanceado Corregido genera fuertes inequidades entre los estudiantes con distintos niveles de desempeño para 3 y 5 opciones también, obteniendo los examinados con mayor nivel de desempeño mayor ventaja porcentual que los de menor nivel de desempeño.

En el modelo Sesgo Céntrico, para la estrategia Pura C, al igual que en 4 opciones, para 3 y 5 opciones no se observó influencia del nivel de desempeño en la ventaja porcentual, pero si se observó influencia en el caso de la desventaja de la estrategia Underdog, donde la desventaja porcentual aumentaba mientras mayor era el nivel de desempeño del examinado. Esto último puede deberse a que mientras mayor sea el número de ítems respondidos por conocimiento, más se acerca la frecuencia de cada opción en estos ítems a la frecuencia dada por el sesgo céntrico (por la ley de los grandes números), y con ello la desventaja se acerca a la desventaja de escoger la opción con menor frecuencia en el sesgo céntrico. Así, el modelo Sesgo Céntrico también puede mostrar inequidades, pero en este caso, los estudiantes de mayor desempeño tendrán mayor riesgo de perder puntos, que los estudiantes de menor desempeño.

Con respecto al número de ítems, en el modelo Sesgo Céntrico la ganancia porcentual de Pura C no se vió afectada, en cambio la desventaja porcentual de Underdog fue mayor mientras mayor fue el número de ítems. El razonamiento usado para explicar la influencia del nivel de desempeño en la desventaja de Underdog podría también explicar este efecto: Mientras más ítems tenga la prueba, es más probable encontrar distribuciones parecidas a los porcentajes del sesgo para un mismo porcentaje de ítems (por la ley de los grandes números) y con ello, Underdog elegiría la opción con menor frecuencia en el Sesgo Céntrico. Para el modelo Balanceado Corregido, al igual que en 4 opciones, una gran cantidad de ítems parece atenuar un poco el ruido, pero incluso para 100 ítems este efecto sigue siendo comparable al efecto en Sesgo Céntrico. Con estos resultados, ambos modelos no son recomendables ni siquiera para un alto número de ítems, dado que en Balanceado Corregido la cantidad de ítems no regula el ruido, y en Sesgo Céntrico un gran número de ítems, de hecho, hace crecer este ruido.

Respecto al Estudio 2, se replicaron las resultados de ganancias de la estrategia Underdog en los modelos Balanceado Exacto, Balanceado Corregido y Random Corregido, mostrando que, al igual que en 4 opciones, la estrategia Underdog saca provecho principalmente del balanceo global y no mucho de la evitación de repeticiones. También se replicaron las ganancias de Underdog en los modelos balanceados aproximativamente, mostrando que, incluso cuando el δ fue mucho más grande para 3 y 5 opciones que para el estudio en 4 opciones, Underdog seguía sacando una ventaja no despreciable. También fue replicada la ganancia de la estrategia Underdog mejorado (incluyendo que fue mayor que Underdog en Balanceado Corregido), indicando que la amenaza a la validez de los datos de este modelo podría ser aún peor que con Underdog en 3 y 5 opciones también. Por último, la ganancia de la estrategia

Evitar Repeticiones Dobles en el modelo Random Corregido también fue replicada para los otros números de opciones.

El tamaño de efecto en las ventajas tuvo 3 comportamientos, dependiendo de que característica del modelo la estrategia sacaba ventaja: 1) si la estrategia sacaba ventaja del balanceo (Underdog Mejorado en Balanceado Exacto y Underdog en Balanceo Exacto), entonces los tamaños de efecto fueron más grandes para un número de opciones más grande, 2) si la estrategia sacaba ventaja de la evitación de repeticiones (Evitar Repeticiones Dobles en Balanceado Corregido y Random Corregido, y Underdog Mejorado en Random Corregido), entonces los tamaños de efecto fueron más grandes para un número de opciones más pequeño y 3) si la estrategia sacaba ventaja del balanceo y de la evitación de repeticiones (Underdog Mejorado en Balanceado Corregido) los tamaños de efecto no tenían una dirección clara.

Un hallazgo importante para la replicación del Estudio 2 es que en 3 opciones el modelo Random Corregido permitió una ventaja con el uso de la estrategia Evitar Repeticiones Dobles (efecto que es muy pequeño en 4 opciones y que no está en 5 opciones), mostrando nuevamente que usar un bajo número de opciones puede conllevar riesgos, y que la recomendación de evitar repeticiones no pareciera adecuada.

Con respecto al rango del δ en 3 y 5 opciones, podemos notar que aunque el δ máximo (14) fue casi el doble del delta del Estudio 4 (8), Underdog siguió teniendo una ventaja no despreciable. Para hacerse una idea de la magnitud de este cambio en el rango de δ en 100 ítems, un modelo aproximativo con $\delta = 8$ puede permitir 2 opciones con una cantidad de 21 y 29 claves cada una (para 4 opciones), 29 y 37 claves cada una (para 3 opciones), 16 y 24 claves cada una (para 5 opciones). Por el otro lado, un $\delta = 14$ puede permitir 2 opciones con una cantidad de 26 y 40 claves cada una (para 3 opciones), 13 y 27 claves cada una (para 5 opciones). Pese a que la magnitud en que difieren el número de claves por opción podría parecer bastante grande, no es suficiente para evitar la ganancia de la estrategia Underdog en ningún número de opciones.

Para el nivel de desempeño, los efectos del Estudio 2 fueron replicados: 1) en todas las ganancias de Underdog (incluyendo en los modelos Balanceado Aproximativo) la ventaja porcentual aumentaba si aumentaba el nivel de desempeño, 2) asimismo, la ventaja porcentual aumentaba cuando aumentaba el nivel para la ganancia de Underdog Mejorado en los modelos Balanceado Exacto y Balanceado Corregido. Para 3 opciones, las ganancias porcentuales de Underdog Mejorado y Evitar Repeticiones Dobles también aumentaron con respecto al nivel de desempeño, fenómeno que no fue observado ni en 4 ni en 5 opciones. Así, las estrategias también generan inequidades en los modelos para 3 y 5 opciones, e incluso más inequidades en 3 opciones.

Se replicó el efecto del número de ítems en el Estudio 2 de la estrategia Underdog en todas las ganancias (incluyendo los modelos Balanceado Aproximativo): Mientras menor era el número de ítems, mayor fue la ventaja porcentual. Este mismo efecto fue notado, al igual que en 4 opciones, para la estrategia Underdog Mejorado en los modelos Balanceado Corregido y Balanceado Exacto. Además, también se replicó que, para un bajo número de ítems, la ventaja que permitían los modelos Balanceado Aproximativo era igual que la ventaja que permitía Balanceado Exacto, lo que muestra la poca efectividad del balanceo aproximado para evitar el ruido generado por estrategias basadas en la posición, para 3 y 5

opciones también.

Respecto a la replicación del Estudio 3, los efectos encontrados en el modelo Random Corregido Por Repeticiones Simples y Random Corregido Por Repeticiones Dobles fueron encontrados también para 3 y 5 opciones.

Para 3 opciones, hubo una nueva ventaja, y es que la estrategia Evitar Repeticiones Simples sacó una ventaja en el modelo Random Corregido Por Repeticiones Triples (con tamaño de efecto de 0.3 al menos).

Se puede entender el efecto nuevo aparecido en Random Corregido Por Repeticiones Triples para 3 opciones y el efecto desaparecido en Random Corregido Por Repeticiones dobles para 5 opciones observando que las estrategias que sacan provecho de la evitación de repeticiones tuvieron mayor tamaño de efecto en su ventaja en número de opciones bajo que alto.

Con respecto a la influencia del nivel de desempeño, se puede notar que todas las ventajas porcentuales que se encontraron son crecientes con respecto al nivel de desempeño, aunque la pendiente se nota más cuando la ventaja era grande, como en el caso de Evitar Repeticiones Simples en Random Corregido Por Repeticiones Simples. Por otra parte, el número ítems no tuvo ninguna influencia.

Los modelos que evitan repeticiones ponen en riesgo la validez de los resultados y además crean inequidades para: 1) repeticiones simples, dobles y triples en el caso de 3 opciones y 2) repeticiones simples y dobles en el caso de 4 y 5 opciones. Así, para pruebas de 3 opciones no es recomendable evitar ningún tipo de repetición, y para pruebas de 4 o 5 opciones no es recomendable evitar repeticiones simples ni dobles. Simplificando, pareciera que evitar repetición no es una característica deseable en un protocolo de posicionamiento de claves.

Si se toma en cuenta los tamaños de efecto de todas las ganancias de los modelos basadas en la recomendación balancear y/o evitar repeticiones (de los Estudios 1, 2 y 3), se puede entender los tamaños de efecto de la siguiente manera: 1) si la ventaja viene de que la estrategia saca ventaja de que el modelo balancea sus claves (Underdog en Balanceado Exacto o Balanceado Corregido; Underdog Mejorado en Balanceado Exacto), entonces los efectos son más grandes en 5 que en 4 opciones, y más grandes en 4 que en 3 opciones y 2) si la ventaja viene de que la estrategia saca ventaja de que el modelo evita repeticiones (Evitar Repeticiones Simples, Dobles o Triples en Random Corregido Por Repeticiones Simples, Dobles o Triples; Evitar Repeticiones Dobles en Balanceado Corregido; Underdog Mejorado en Random Corregido Por Repeticiones Dobles), entonces los efectos son más grandes en 3 que en 4 opciones, y más grandes en 4 que en 5 opciones.

Capítulo 6

Discusión general

En este trabajo, se ha podido demostrar que el uso de estrategias basadas en la posición de opciones correctas por parte de examinados puede teóricamente modificar los puntajes obtenidos en pruebas de selección múltiple de manera significativa y, en algunos casos, preocupante. Estas estrategias pueden tener influencia en todos los niveles de desempeño, y aunque las pruebas tengan diferentes números de ítems y números de opciones. Diferentes estrategias se asocian a diferentes ganancias de puntajes, según el posicionamiento de claves realizado por los constructores, pero la mayoría de los métodos de posicionamiento de claves, basados en recomendaciones usuales de guías de construcción de pruebas, están sujetos a que los examinados puedan potencialmente aprovecharlos gracias a estas estrategias.

6.1. Problemas de validez e inequidad

En pruebas cuya posición de opciones correctas no fue controlada en el ensamblaje (modelo basado en el sesgo céntrico), ciertas estrategias de resolución (Pura C o Pura B) pueden sacar ventaja, tengan las pruebas 3, 4, o 5 opciones. Por tanto, se podría decir que preocuparse por el posicionamiento de las claves en pruebas de selección múltiple es relevante: al no cuidar este factor, se corre el riesgo de que el uso de estrategias por los estudiantes haga perder validez a las mediciones.

Preocuparse por el posicionamiento de las opciones correctas no garantiza resolver el problema. En todos los modelos que siguen la recomendación de balanceo global (Balanceo Exacto, Balanceo Aproximativo, y Balanceo Corregido), la estrategia Underdog pudo sacar una ventaja significativa y no despreciable, incluso en los modelos aproximativos, que podrían haber sido pensados como modelos muy parecidos al modelo Random. En los modelos basados en la recomendación de balanceo local (evitar repeticiones), también se encontraron ventajas, esta vez de la estrategia de evitación de repeticiones, que se trate de repeticiones simples (en 3, 4, y 5 opciones), dobles (en 3 y 4 opciones), e incluso triples (en 3 opciones).

Además, los métodos de balanceo de claves en pruebas no solamente conllevan potenciales problemas de validez, sino también pueden conllevar problemas de equidad. El nivel de des-

empeño afectó a la ventaja en todos los modelos de balanceo, siendo la ventaja creciente con respecto al nivel. El nivel de desempeño también afectó a todas las ventajas en los modelos de evitación de repeticiones, siendo también crecientes con respecto al nivel, pero este efecto fue notorio sólo cuando la ventaja tuvo tamaño de efecto por lo menos pequeño. Por lo tanto, tampoco es recomendable posicionar la clave de una prueba siguiendo la recomendación del balanceo en cualquiera de sus formas o la de evitar repeticiones, debido a que el uso de estrategias por parte de los examinados, además de poner en riesgo la validez de los resultados, genera también inequidad en los resultados.

El uso de estrategias no solamente puede traer ganancias de puntos, sino que también puede hacer perder puntos. Por ejemplo, en el modelo Sesgo Céntrico se encontró una desventaja en el uso de la estrategia Underdog, y esta desventaja fue mayor mientras mayor fue el nivel de desempeño de los examinados y mayor fue el número de ítems. Por tanto, el uso de estrategias podría ser perjudicial para un examinado, dado que en la mayoría de los casos un examinado se enfrenta a una prueba sin saber como fue atribuida la posición de las claves, y esto podría ser terrible en el caso de las pruebas de alta consecuencia que en general tienen una alta cantidad de ítems. Se recomienda, entonces a quienes administran las pruebas, al igual que en [32], recalcar este aspecto a los examinados, para invitarlos a no usar estrategias porque podrían desfavorecerlos.

En todos los modelos de balanceo, la ventaja fue creciente mientras menor fue el número de ítems, siendo este efecto muy pronunciado y por ello preocupante. Además, cuando el número de ítems fue bajo, la atenuación de la ventaja que tenía usar un balanceo aproximativo en vez de un balanceo exacto desaparece, por lo en el caso de balancear es altamente no recomendable usar un número bajo de ítems. En los casos del modelo Sesgo Céntrico y los modelos de evitación de repeticiones, no obstante, el número de ítems no tiene ninguna incidencia.

En el modelo Sesgo Céntrico, los tamaños de efecto de la ventaja de Pura C o Pura B fueron más grandes en 3 que 4 y en 4 que 5 opciones, esto debido a que el efecto sesgo céntrico fue más pronunciado en 3 que en 4 que en 5 opciones. Las magnitudes del sesgo céntrico podrían explicarse como una preferencia de las opciones centrales, como en 5 opciones hay 3 opciones centrales (B,C y D), entonces la preferencia con respecto a las opciones extremales (A y E) es menos fuerte que en el caso en que solo hay 2 opciones centrales (4 opciones: B y C), o en el que sólo hay una opción central (3 opciones: C). La ventaja de la estrategia Underdog en los modelos de balanceo tuvo un tamaño de efecto sólo un poco más grande en 5 que en 4 opciones, y en 4 que en 3 opciones, pero cuando la estrategia se basó en evitar repeticiones (en los modelos Random Corregido o el modelo Balanceado Corregido), los tamaños de efecto fueron mucho más grandes en 3 que en 4 opciones y en 4 que en 5 opciones. En general entonces, los efectos de las estrategias basadas en la posición de las opciones correctas sufren un fuerte incremento cuando hay menos opciones, cuestionando la recomendación de usar ítems con 3 opciones [34], basado en el supuesto que este número de opciones no tendría efectos perjudiciales sobre la calidad psicométrica de las puntuaciones en pruebas de selección múltiple. En este trajo se han propuesto varias estrategias que pueden sacar ventajas de ciertas recomendaciones de posicionamiento de claves, no obstante, la ventaja que podrían sacar otras estrategias podría ser mayor a las que hemos presentado. Por ejemplo, en el caso del modelo Balanceado Corregido, se pudo construir una estrategia un poco mejor que Underdog (Underdog Mejorado), y en el caso de Random Corregido Por Repeticiones Dobles, también se pudo construir una estrategia mejor (Evitar Repeticiones Simples) que la

que primero se propuso (Evitar Repeticiones Dobles). Sin contar los modelos en los que se puede construir explícitamente la estrategia de ganancia maximal (Underdog en Balanceado Exacto, escoger la opción con mayor frecuencia en Sesgo Céntrico, y cualquier estrategia en Random porque todas las estrategias son maximales), para el resto de modelos no se ha podido construir esta estrategia (cuya existencia fue probada teóricamente). Por tanto, es posible que existan estrategias que tengan mayor ganancia que las presentadas, y con ello, que tanto el riesgo de validez como la inequidad sean mayores a las aquí presentadas.

Para todas las estrategias de resolución de este estudio, y más generalmente para todas las estrategias basadas en la posición de las opciones, el modelo basado en la directriz randomizar la posición de la clave"no permite ganancias diferentes a las que se obtienen con la estrategia Random. Por ello, este modelo no presenta ni problemas de validez ni problemas de equidad. Esto es cierto para cualquier número de ítems, cualquier número de opciones, y cualquier nivel de desempeño (ver Corolario A.8). Es recomendable entonces seguir la directriz de randomizar la posición de las claves en la construcción de pruebas de selección múltiple.

6.2. Las recomendaciones de balanceo no están obsoletas

Una revisión sistemática de recomendaciones sobre posicionamiento de claves muestra que las recomendaciones de balanceo global y local no están obsoletas [24], entonces uno podría preguntarse, ¿por qué se siguen dando estas recomendaciones pese al llamado a no usarlas [1, 3]? La respuesta podría tener relación con el llamado "juicio de representatividad random", que es la expectativa que tiene un estudiante (o un profesor) de cómo debería ser una distribución de claves que fue generada desde una aleatoriedad random. En [23], se muestra que cuando estudiantes tenían que adivinar en que opción (A ó B) estaba la alternativa correcta de un ítems mostrado en una hoja de respuestas, ítem que seguía de una racha 3 o 4 claves seguidas, entonces elegían más frecuentemente la clave distinta de la racha, apoyando así que patrones como las rachas son menos "representativos"de una prueba generada random, y por ello, hay menor expectativa de que sean la respuesta de una prueba. Este rechazo a esperar patrones en las respuestas también fue mostrado en [32], donde se mostró que en una prueba real, los patrones del tipo AAAABBBB...DDDD y el tipo ABCDABCD...ABCD en la distribución de las claves afectaban negativamente el puntaje de los examinados, y de hecho, el efecto era peor para los examinados con mayor puntaje (posiblemente porque eran quienes más podían ver el patrón). Podemos suponer que el juicio de representatividad podría influir en los constructores (al igual que en los estudiantes) al decidir si aceptar o no una distribución de claves generadas aleatoriamente, y por ello, las recomendaciones de balanceo global y local seguirían vigentes.

Aunque podemos decir que los patrones largos afectan el rendimiento de los estudiantes, lo cierto es que la probabilidad de aparición de estos patrones cuando una prueba es generada aleatoriamente es muy baja, por lo que en la práctica no tendría un efecto importante. Otros estudios han buscado efectos en el rendimiento de estudiantes que se enfrentan a pruebas construidas con recomendaciones más recurrentes, por ejemplo, en [21], el rendimiento de los examinados frente una prueba con rachas de 3 o 4 claves seguidas fue comparado con el

rendimiento frente a una prueba con claves alternadas, pero ningún efecto en el tratamiento se observó, sólo fue observado que cuando los examinados se enfrentaban a responder después de 2 claves seguidas, fue más probable que evitaran dar la misma respuesta que en el ítems anterior, aunque para el caso de una clave o tres claves, fue más probable que dieran la misma respuesta que el ítem anterior, por lo que estos resultados son más bien inconcluyentes. Por otra parte, en [20], el rendimiento entre examinados que realizaban pruebas generadas random y pruebas aproximadamente balanceadas también fue testeado, pero nuevamente ningún efecto fue encontrado. Pareciera entonces que sólo los patrones muy largos y evidentes podrían afectar las respuestas de los estudiantes, y que las pruebas que siguen recomendaciones de balanceo global o balanceo local no afectan las las respuestas de los estudiantes. Es importante destacar, que estos experimentos no son en pruebas de alta consecuencia, por lo que los estudiantes a priori podrían no tener incentivos para usar estrategias basadas en la posición de las opciones, y por eso, los rendimientos en distintos tipos de pruebas podrían ser los mismos a menos que fueran casos muy extremos como los patrones largos. Dado que el balanceo local o global no tiene efectos en las respuestas al igual que random, y que en realidad, la probabilidad de que eviten un patrón muy largo es muy baja (por que en general la probabilidad de encontrar un patrón muy largo es baja), entonces no se justificaría usar estas directrices en vez de random.

Si se quisiera evitar patrones muy evidentes en la construcción de una prueba (como un nuevo modelo de contrucción), la ventaja del uso de estrategias podría también ser estimada. Primero, como la frecuencia de aparición de estos patrones es baja en general en las secuencias random, podría esperarse, a priori, que la ventaja de alguna estrategia basada en estos patrones no pudiera ser significativa. Segundo, para probar que evitar algún tipo de patrón en una prueba no afecta la validez o equidad de los resultados, podría simularse directamente la ganancia de la estrategia de ganancia maximal y la ganancia de la estrategia de ganancia minimal, para acotar el ruido de este modelo, usando las formas explícitas del Teorema A.10 y del Corolario A.11. Pero sin más estudios que demuestren un efecto en el rendimiento cuando hay patrones en la distribución de las claves, no se justificaría corregir una prueba, por lo que la mejor recomendación para la contrucción de pruebas de selección múltiple sigue siendo randomizar.

Cabe mencionar que una estrategia fue definida con una componente estocástica, esto quiere decir que para llevar a cabo esta estrategia en la realidad se necesitaria un dispositivo que genere aleatoriedad, porque en general los humanos son incapaces de generar prácticas aleatorias tales como elegir uniformemente. Esto no es un impedimento para los resultados de los estudios, dado que para cualquier estrategia (que podría o no ser estocástica) hay al menos una estrategia determinista (y por ello replicable en la realidad) que tiene mayor ganancia (ver Teorema A.10). Con esto, las ventajas vistas en este trabajo siempre son alcanzables usando una estrategia determinista, lo que le da validez práctica a los resultados de este trabajo. Cabe mencionar que las estrategias de ganancia maximal podrían no poder aplicarse de forma simple como las que vimos en este trabajo, que se podían aplicar en general siguiendo un par de reglas.

Las implicancias de la existencia de una estrategia de ganancia maximal y una estrategia de ganancia minimal para cada modelo son incluso más todavía: La posible ventaja o desventaja que pueda tener una estrategia en el modelo estará acotada por las ganancias de las estrategias de ganancia maximal y minimal respectivamente. Con ello, si se quisiera entonces

proponer una recomendación para posicionar las claves en una prueba de selección múltiple, la ventaja y desventaja calculadas con ambas estrategias (maximal y minimal) serviría como indicador de la vulnerabilidad de la recomendación ante estrategias basadas en la posición de las opciones correctas.

6.3. Limitaciones

Una limitación importante de este trabajo es que es un estudio que simula conductas, no que las observe. Hay registro de estudios que observan las conductas de posicionamiento de opciones [1, 3], estudios que observan las conductas de los examinados en las respuestas cuando adivinan [1, 8], y también estudios que observan el rendimiento de estudiantes en pruebas construidas con diferentes recomendaciones [21, 20] o en pruebas con distribuciones de claves manipuladas [32], sin embargo, no hay registro de estudios que evalúen la interacción entre la distribución de las claves en una prueba y el uso de conductas estratégicas para adivinar la respuesta correcta. Así, este estudio es el primero que muestra claramente cómo podrían interactuar estas dos conductas, concordando con [2] en que este fenómeno se tiene que abordar como un juego social. Aún así, faltan estudios que evalúen la interacción de estas dos conductas en un entorno experimental.

Otra de las limitaciones de este trabajo es el supuesto de que las preguntas respondidas por conocimiento son siempre correctas: Un examinado muy bien puede responder por conocimiento de forma errada, por realizar un error conceptual, o por inatención, etc. Si se abordara la resolución de los ítems de una prueba con la teoría de respuesta al ítem (IRT), la resolución de cada ítem podría ser modelada como la probabilidad de responder bien un ítem, dado el nivel del examinado y dada la dificultad del ítem. De esta manera, se podrían modelar patrones de respuesta por conocimiento más precisos y auténticos, consiguiendo así un acercamiento más real al proceso de resolución. Se podría además, en un modelo IRT de varios parámetros, considerar un parametro que modelara la probabilidad de adivinar la posición de la opción correcta, lo que permitiría precisar aún más el cálculo de la probabilidad de que una estrategia de resolución basada en la posición de las opciones correctas sea exitosa o no.

Otra limitación de este trabajo es que se ha considerado que el proceso de resolución por estrategia no entra en interacción con el proceso de resolución por conocimiento, ya que estos dos procesos se desencadenan para ítems diferentes. Se ha considerado que cuando se responde por estrategia, se hace sin conocimiento y por tanto se consideran todas las posiciones, aunque podría considerarse que gracias a un conocimiento parcial se podría tener descartadas algunas opciones de antemano. Modelar cada distractor con un indice de descartabilidad (usando su proporción de respuestas en aplicaciones previas, por ejemplo) podría acercar de mejor manera al proceso de resolución por estrategia y su interacción con el conocimiento.

Por último, este trabajo tampoco considera todas las estrategias de posicionamiento de claves documentadas, como por ejemplo, casos de ítems/pruebas en los cuales se ordenan las opciones de forma numerica, alfabética, o por tamaño de la opción. Sin embargo, sin demostración clara de que en estos casos el ordenamiento conduce a una distribución de claves

sesgadas de algún modo, este ordenamiento de opciones debería conducir a distribuciones de claves parecidas a random.

Capítulo 7

Conclusión

En este trabajo, se ha mostrado que no preocuparse por la posición de las opciones correctas en una prueba de selección múltiple pone en riesgo la validez de los resultados de la prueba. Adicionalmente, se ha mostrado que controlar la posición de las opciones correctas siguiendo recomendaciones como balancear global y/o localmente también pone en riesgo la validez de los resultados, además de, en el caso de balancear globalmente, generar problemas de inequidad en los resultados. Estos hallazgos se observan en pruebas con un amplio rango de número de ítems, sean las pruebas de 3, 4, o 5 opciones. Son observables para examinados de todos los niveles de desempeño. Se concluye que randomizar la posición de las opciones correctas es el único método de posicionamiento que no genera problemas de validez ni equidad en los resultados (en las situaciones estudiadas en este trabajo al menos) y que, por tanto, pareciera ser el método más adecuado para cuidar que la distribución de las opciones correctas no proporcione pistas permitiendo a examinados entregar algunas respuestas correctas cuando no cuentan con el conocimiento disciplinar requerido.

A través de los cuatro estudios realizados, este trabajo deja constancia que la realización de un proyecto en el ámbito matemático permite aportar datos de alta relevancia para el diseño de prácticas educativas, en particular en evaluación. También demuestra que un estudio de simulaciones tiene la capacidad de modelizar un fenoméno social, potencialmente dificilmente observable en contexto ecológico. Se ha llamado a constituir una base de evidencias científicas sólidas que respalde las directrices de construcción de pruebas y preguntas de selección múltiple y permita así confirmar que la experiencia acumulada docente no lleva a los constructores a tomar acciones innecesarias o contraproducentes. Este trabajo responde a este llamado y, con ello, espera aportar al diseño de un recopilado de directrices de construcción basadas en la evidencia, que sean garantizadoras de que los instrumentos de evaluación del futuro permitan medir de forma aún más limpia y eficiente el aprendizaje.

Bibliografía

[1] Yigal Attali and Maya Bar-Hillel. Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, 40(2):109–128, 2003.

[2] Maya Bar-Hillel. Position effects in choice from simultaneous displays: A conundrum solved. *Perspectives on Psychological Science*, 10(4):419–433, 2015.

[3] Maya Bar-Hillel and Yigal Attali. Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician*, 56(4):299–303, 2002.

[4] Don F Blood and William C Budd. *Educational measurement and evaluation*. Harper and Row, 1972.

[5] Mybert Eustace Broom. *Educational measurements in the elementary school*. McGraw-Hill Book Company, London, 1939.

[6] George A Brown, Joanna Bull, and Malcolm Pendlebury. *Assessing student learning in higher education*. Routledge, 2013.

[7] Jacqueline A Carnegie. Does correct answer distribution influence student choices when writing multiple choice examinations?. *Canadian Journal for the Scholarship of Teaching and Learning*, 8(1):11, 2017.

[8] Kathy Carter. Test-wiseness for teachers and students. *Educational Measurement: Issues and Practice*, 5(4):20–23, 1986.

[9] Seth DeVore, John Stewart, and Gay Stewart. Examining the effects of testwiseness in conceptual physics evaluations. *Physical Review Physics Education Research*, 12(2):020138, 2016.

[10] Carol A Dwyer. Achievement testing. *Encyclopedia of educational research*, 1, 1982.

[11] Randy A Ellsworth, Pat Dunnell, and Orpha K Duell. Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, 83(5):289–293, 1990.

[12] Max D Engelhart. Suggestions for writing achievement exercises to be used in tests scored on the electric scoring machine. *Educational and Psychological Measurement*, 7(3):357–374, 1947.

[13] Marshall A Geiger and Mark M Higgins. On the answer-arrangement bias of professional certification examinations in accounting. *Accounting Educators' Journal*, 9(2):89, 1997.

[14] Mark J Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6):1082–1116, 2017.

[15] Thomas M Haladyna. Developing and validating multiple-choice test items. 2004.

[16] Thomas M Haladyna and Steven M Downing. A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, 2(1):37–50, 1989.

[17] Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.

[18] TM Haladyna and MC Rodriguez. Guidelines for writing selected-response items. *Developing and validating test items*, pages 89–110, 2013.

[19] Nicole L Ibbett and Brett J Wheldon. The incidence of clueing in multiple choice testbank questions in accounting: Some evidence from australia. *e-Journal of Business Education and Scholarship of Teaching*, 10(1):20–35, 2016.

[20] Dane Christian Joseph. Randomize it: fair procedures when constructing multiple-choice test-keys. *Journal of Effective Teaching in Higher Education*, 2(1):80–92, 2019.

[21] Hubert János Kiss and Adrienn Selei. Do streaks matter in multiple-choice tests? *Education Economics*, 26(2):179–193, 2018.

[22] EB Kolawole. A survey of the anchor bias in mathematics objective tests in west african examination council (waec). *Journal of Emerging Trends in Educational Research and Policy Studies*, 2(3):171–173, 2011.

[23] Chan Jean Lee. The test taker's fallacy: How students guess answers on multiple-choice tests. *Journal of Behavioral Decision Making*, 32(2):140–151, 2019.

[24] Séverin Lions, María Paz Blanco, Pablo Dartnell, Carlos Monsalve, Gabriel Ortega, and Julie Lemarié. Item-writing guidelines on response options placement: A systematic review. *Educational Research*, 2022. Under review.

[25] Séverin Lions, Carlos Monsalve, Pablo Dartnell, María Paz Blanco, Gabriel Ortega, and Julie Lemarié. Does the response options placement provide clues to the correct answers in multiple-choice tests? a systematic review. *Applied Measurement in Education*, pages 1–20, 2022.

[26] Joan C Masters, Barbara S Hulsmeyer, Mary E Pike, Kathy Leichty, Margaret T Miller, and Amy L Verst. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education, 2001.

[27] Mary E McDonald. *The nurse educator's guide to assessing learning outcomes*. Jones & Bartlett Learning, 2017.

[28] N. S Metfessel. An experimental analysis of response sets in forced choice test performance. *Unpublished doctoral dissertation, University of Southern California*, 1955.

[29] Newton S Metfessel and Gilbert Sax. Systematic biases in the keying of correct responses on certain standardized tests. *Educational and Psychological Measurement*, 18(4):787–790, 1958.

[30] William E Mitchell. Bias in writing objective-type examination questions. *Journal of Economic Education*, pages 58–60, 1974.

[31] Martin Johan Nelson. *Tests and Measurements in Elementary Education*. Cordon Company, 1939.

[32] Stephen T Paul, Samantha Monda, S Maria Olausson, and Brenna Reed-Daley. Effects of apophenia on multiple-choice exam performance. *SAGE Open*, 4(4):2158244014556628, 2014.

[33] Henry Daniel Rinsland. Constructing tests and grading in elementary and high school subjects. 1937.

[34] Michael C Rodriguez. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational measurement: issues and practice*, 24(2):3–13, 2005.

[35] Michael C Rodriguez and Anthony D Albano. *The college instructor's guide to writing test items: Measuring student learning*. Routledge, 2017.

[36] María Cristina Rodríguez-Díez, Manuel Alegre, Nieves Díez, Leire Arbea, and Marta Ferrer. Technical flaws in multiple-choice questions in the access exam to medical specialties ("examen mir") in spain (2009–2013). *BMC medical education*, 16(1):1–8, 2016.

[37] Giles Murrel Ruch. *The objective or new-type examination: An introduction to educational measurement*. Scott, Foresman, 1929.

[38] Bonnie R Rush, David C Rankin, and Brad J White. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*, 16(1):1–10, 2016.

[39] Marie Tarrant, Aimee Knierim, Sasha K Hayes, and James Ware. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26(8):662–671, 2006.

[40] James B Trump and Helen R Haggerty. *Basic principles in achievement test item construction*. Personnel Research Section, Personnel Research and Procedures Branch ..., 1965.

Anexos

Anexo A

Medida de máxima entropía y estrategia de ganancia maximal

En este capítulo se darán las principales definiciones usadas a lo largo del texto. Primero definiremos la entropía de un espacio de medida, que es manera de medir la complejidad de este espacio, que puede ser visto como una fuente de información. Luego calcularemos medidas que maximizan la entropía de ciertos sistemas. Por último, definiremos estrategia y ganancia para mostrar la existencia de una estrategia de ganancia maximal.

A.1. Medidas de máxima entropía

En adelante trabajaremos con espacios discretos, por lo que en adelante X será un espacio discreto y finito.

Definición A.1 (Entropía) *Sea $(X, \mathcal{F}, \mu)$ un espacio de medida finita, se define la entropía del espacio como:*

$$H(\mu) := \sum_{x \in X : \mu(x) > 0} -\mu(x) \log(\mu(x))$$

se escribirá $H(X, \mathcal{F}, \mu) := H(\mu)$ cuando se quiera especificar el sistema.

Si fijamos $\mathcal{F}$ como $\mathcal{P}(X)$ el conjunto de las partes de X, y tomamos el conjunto $\mathcal{M}_K(X, \mathcal{F})$ de las medidas definidas en $(X, \mathcal{F})$ tal que $\mu(X) = K$, podemos calcular la medida que maximiza la entropía del sistema:

Proposición A.2 *Con la notación anterior:*

$$\arg\max_{\mu \in \mathcal{M}_K(X,\mathcal{F})} H(\mu) = \{K \cdot Unif(X)\}$$

con $Unif(X)$ la medida de probabilidad uniforme en X. Además, si tenemos un subconjunto $A \subset X$, también podemos calcular el máximo en las medidas concentradas en A:

$$\arg\max_{\mu \in \mathcal{M}_K(X,\mathcal{F}) \wedge \mu(A^c)=0} H(\mu) = \{K \cdot Unif(A)\}$$

Corolario A.3 *Fijando X como el conjunto de secuencias $\mathcal{A}^N$, con $\mathcal{A}$ finito, podemos saber la forma de casi todos los modelos de construcción (definidos como sistemas de máxima entropía) del texto:*

- *Modelo Random:*
  $\rightarrow Unif(\mathcal{A}^N)$
- *Modelo Balanceado Corregido:*
  $\rightarrow Unif(\{Secuencias\ balanceadas\ sin\ repeticiones\ dobles\})$
- *Modelo Balanceado:*
  $\rightarrow Unif(\{Secuencias\ balanceadas\})$
- *Modelo Balanceado Aproximativo δ:*
  $\rightarrow Unif(\{Secuencias\ balanceadas\ aproximadamente\ con\ el\ criterio\ delta\})$
- *Modelo Random Corregido Por Repeticiones Simples:*
  $\rightarrow Unif(\{Secuencias\ sin\ repeticiones\ simples\})$
- *Modelo Random Corregido Por Repeticiones Dobles:*
  $\rightarrow Unif(\{Secuencias\ sin\ repeticiones\ dobles\})$
- *Modelo Random Corregido Por Repeticiones Triples:*
  $\rightarrow Unif(\{Secuencias\ sin\ repeticiones\ triples\})$

Otra propiedad importante de la entropía es la relación de la entropía de un sistema con la entropía de los subsistemas que lo componen:

Proposición A.4 *Sea $X = \prod_{1\leq i\leq N} X^i$ un espacio producto, $(X, \mathcal{P}(X), \mu)$ un espacio de probabilidad, si definimos*

$$\mu_i : \mathcal{P}(X^i) \rightarrow \mathbb{R}; \mu_i(A) = \mu(X^1, ..., X^{i-1}, A, X^{i-1}, ..., X^N),$$

entonces tenemos que:

$$H(X, \mathcal{P}(X), \mu) \leq H\left(X, \mathcal{P}(X), \bigotimes_{1\leq i\leq N} \mu_i\right) \leq \sum_{1\leq i\leq N} H(X^i, \mathcal{P}(X^i), \mu_i)$$

Además, si $\mu \neq \bigotimes_{1\leq i\leq N} \mu_i$, entonces $H(X, \mathcal{P}(X), \mu) < H(X, \mathcal{P}(X), \bigotimes_{1\leq i\leq N} \mu_i)$

Esta propiedad es útil para calcular la medida de máxima entropía en el caso del sesgo céntrico:

Corolario A.5 *Sea $\mathcal{M}_1(\mathcal{A}^N, \mathcal{P}(A))$ el espacio de las medidas de probabilidad en $(\mathcal{A}^N, \mathcal{P}(A))$. Sea también N_{alt} la función que cuenta la cantidad de elementos alt de la secuencia, i.e.:*

$$N_{alt} : \mathcal{A}^N \rightarrow \mathbb{N}; N_{alt}(p) = \#\{i \in \{1, .., N\} : p_i = alt\}$$

si definimos:

$$\mathcal{K} = \{\mu \in \mathcal{M}_1(\mathcal{A}^N, \mathcal{P}(\mathcal{A})) : \mathbb{E}_\mu\left(\frac{N_{alt}}{N}\right) = p_{alt}, \forall alt \in \mathcal{A}\}$$

con $(p_{alt})_{alt\in\mathcal{A}} \subset \mathbb{R}^+$ *una secuencia de pesos estrictamente positivos de una probabilidad discreta, entonces:*

$$\arg\max_{\mu\in\mathcal{K}} H(\mu) = \left\{ \bigotimes_{1\leq i\leq N} \mathbb{P}_{disc}((p_{alt})_{alt\in\mathcal{A}}) \right\}$$

con $\mathbb{P}_{disc}((p_{alt})_{alt\in\mathcal{A}}$ *la probabilidad discreta dada por los pesos* $(p_{alt})_{alt\in\mathcal{A}}$.

Demostración. Primero veremos que si $\mu \in \mathcal{K}$, entonces, si definimos $\mu^* = \bigotimes_{1\leq i\leq N} \mu_i$ (como en la Proposición A.2), entonces $\mu^* \in \mathcal{K}$.
Notamos que $\mu^* \in \mathcal{M}_1(\mathcal{A}^N, \mathcal{P}(\mathcal{A}))$, luego vemos que podemos escribir N_{alt} como:

$$N_{alt} = \sum_{i\leq N} 1_{alt} \circ P_i$$

Así, como $\mu \in \mathcal{K}$ tenemos que:

$$\begin{aligned}
& \forall alt \in \mathcal{A}, \quad \mathbb{E}_\mu\left(\frac{N_{alt}}{N}\right) = p_{alt} \\
\implies & \forall alt \in \mathcal{A}, \quad \mathbb{E}_\mu\left(\sum_{i\leq N} 1_{alt} \circ P_i\right) = p_{alt} N \\
\implies & \forall alt \in \mathcal{A}, \quad \sum_{i\leq N} \mathbb{E}_\mu\left(1_{alt} \circ P_i\right) = p_{alt} N \quad \text{(solo depende la coordenada i)} \\
\implies & \forall alt \in \mathcal{A}, \quad \sum_{i\leq N} \mathbb{E}_{\mu^*}\left(1_{alt} \circ P_i\right) = p_{alt} N \\
\implies & \forall alt \in \mathcal{A}, \quad \sum_{i\leq N} \mathbb{E}_{\mu^*}\left(1_{alt} \circ P_i\right) = p_{alt} N \\
\implies & \forall alt \in \mathcal{A}, \quad \mathbb{E}_{\mu^*}\left(\frac{N_{alt}}{N}\right) = p_{alt} \\
\implies & \mu^* \in \mathcal{K}
\end{aligned}$$

Ahora, por la Proposición A.2 tenemos que:

$$H(\mu) \leq H(\mu^*) \wedge (\mu \neq \mu^* \implies H(\mu) < H(\mu^*))$$

entonces:

$$\arg\max_{\mu\in\mathcal{K}} H(\mu) = \arg\max_{\mu\in\mathcal{K}:\mu=\bigotimes_{1\leq i\leq N}\mu_i} H(\mu)$$

si $\mu = \bigotimes_{1\leq i\leq N} \mu_i$ entonces:

$$\begin{aligned}
\mu \in \mathcal{K} &\iff \sum_{i\leq N} \mathbb{E}_{\mu}(1_{alt} \circ P_i) = p_{alt} N \\
&\iff \sum_{i\leq N} \mathbb{E}_{\mu_i}(1_{alt}) = p_{alt} N \\
&\iff \sum_{i\leq N} \mu_i(\{alt\}) = p_{alt} N \\
&\iff \sum_{i\leq N} \mu_i^{alt} = p_{alt} N \quad (\mu_i^{alt} := \mu_i(\{alt\}))
\end{aligned}$$

Definamos ahora la siguiente medida μ^* de probabilidad:

$$\mu^* := \bigotimes_{1\leq i\leq N} \mu_i^*$$

con $\mu_i^*(alt) := p_{alt}, \forall alt \in \mathcal{A}, \forall i \leq N$. Ahora veamos que $H(\mu) \leq H(\mu^*)$:

$$\begin{aligned}
H(\mu) &= \sum_{i\leq N} H(\mu_i) \\
&= \sum_{i\leq N} \sum_{alt\in\mathcal{A}} -\mu_i^{alt} \log \mu_i^{alt} \\
&= \sum_{alt\in\mathcal{A}} \sum_{i\leq N} -\mu_i^{alt} \log \mu_i^{alt} \\
&= \sum_{alt\in\mathcal{A}} H((\mu_i^{alt})_{i\leq N})
\end{aligned}$$

como $(\mu_i^{alt})_{i\leq N}$ es una medida en un espacio de N elementos, y la medida de este espacio es $\sum_{i\leq N} \mu_i^{alt} = p_{alt} N$, entonces su entropía es menor a la medida uniforme en este espacio, i.e., $(p_{alt})_{i\leq N} = ((\mu_i^*)^{alt})_{i\leq N}$, así:

$$\begin{aligned}
H(\mu) &= \sum_{alt\in\mathcal{A}} H((\mu_i^{alt})_{i\leq N}) \\
&\leq \sum_{alt\in\mathcal{A}} H(((\mu_i^*)^{alt})_{i\leq N}) \\
&= H(\mu^*)
\end{aligned}$$

Así, tenemos que μ^* es un máximo factible del problema. Como el problema de optimización tiene un conjunto de restricciones convexo, y la función de entropía es estrictamente cóncava, entonces el máximo es único, y con esto:

$$\arg\max_{\mu\in\mathcal{K}} H(\mu) = \{\mu^*\}$$

Terminando la demostración. □

A.2. Estrategias de ganancia maximal

Ahora definiremos que será una estrategia. Sea $\mathcal{A}^N$ el espacio de secuencias, y $\emptyset \subsetneq I_{conoc} \subsetneq \{1, ..., N\}$ un conjunto de índices. Supongamos que se tiene acceso a los elementos de una secuencia en los índices de I_{conoc}, el objetivo es adivinar que elementos hay en los índices que no están en I_{conoc}. Una estrategia en I_{conoc} será entonces una función (posiblemente estocástica) que toma los valores de la secuencia en I_{conoc} y entrega posibles valores para la secuencia en $I_{est} := \{1, ..., N\} \setminus I_{conoc}$. Definiendo una función para cada conjunto de índices, queda una definición para una estrategia:

Definición A.6 (Estrategia) *Una estrategia es una familia de funciones F indexada en $\mathcal{P}(\mathcal{A}^N) \setminus \{\emptyset, \mathcal{A}^N\}$, tal que cada elemento F_I es una función:*

$$F_I : \mathcal{A}^I \times \Omega \to \mathcal{A}^{\{1,...,N\}\setminus I}$$

También diremos que la estrategia es determinista si cada elemento de la estrategia F se puede escribir como una función

$$g : \mathcal{A}^I \to \mathcal{A}^{\{1,...,N\}\setminus I}$$

tal que:

$$F_I(p, \omega) = g(p), \forall \omega \in \Omega$$

Siguiendo con la lógica de la adivinanza, también se debe tener una métrica que nos diga que tan buena es. Para el caso de este texto, la ganancia será la cantidad de elementos que adivinamos correctamente. Podemos definir la ganancia de usar una estrategia como:

Definición A.7 (Ganancia) *Sea $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P})$ un espacio de probabilidad, y F una estrategia en $\mathcal{A}^N$. Para un conjunto de índices $\emptyset \subsetneq I \subsetneq \{1, ..., N\}$, y para una prueba $p \in \mathcal{A}^N$, se define la función ganancia en los índices I_{conoc} de la estrategia F como:*

$$G_F(I) : \mathcal{A}^N \times \Omega \to \mathbb{N}; \quad G_F(I)(p, \omega) \longmapsto \sum_{i \in I^c} 1_{\{F_I(P_I(p), \omega)_i = p_i\}}$$

Se define entonces la ganancia promedio para un conjunto de índices $\emptyset \subsetneq I \subsetneq \{1, ..., N\}$ como:

$$G_F(I) \longmapsto \sum_{p \in \mathcal{A}^N} \mathbb{P}(p) \mathbb{E}_\omega \left(\sum_{i \in I^c} 1_{\{w : F_I(P_I(p), \omega)_i = p_i\}} \right)$$

Con esto, decimos que dos estrategias tienen igual ganancia si las distribuciones en $\mathbb{N}$ generadas por sus funciones de ganancia son iguales para todos los conjuntos de índices posibles. También decimos que una estrategia tiene menor ganancia promedio que otra si para cada conjunto de índices, la ganancia promedio es menor.

Es posible demostrar que las ganancias de cualquier estrategia en el modelo Random son iguales, y también que la ganancia de la estrategia Random es igual para todos los modelos, lo que nos entrega un control tanto para los modelos como para las estrategias:

Corolario A.8 (Modelo control) *Sea* $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P} = \bigotimes_{1\leq i\leq N} Unif(\mathcal{A}))$ *el espacio de probabilidad dado por el modelo Random, entonces todas las estrategias definidas en este espacio tienen igual ganancia.*

Demostración. Sea F una estrategia en este espacio y $\emptyset \subsetneq I \subsetneq \{1, ..., N\}$ un conjunto de índices. Primero mostraremos que si $i, j \in I^c$, y $i \neq j$, entonces $1_{\{F_I(P_I(p),\omega)_i = p_i\}}$ y $1_{\{F_I(P_I(p),\omega)_j = p_j\}}$ son independientes. Esta es independencia es equivalente a probar que los conjuntos $B_i := \{F_I(P_I(p), \omega)_i = p_i\}$ y $B_j := \{F_I(P_I(p), \omega)_j = p_j\}$ son independientes. Por un lado tenemos que:

$$\begin{aligned}
\mathbb{P}(B_i \cap B_j) &= \mathbb{P}(\{(p,w) : F_I(P_I,w)_i = p_i\} \cap \{(p,w) : F_I(P_I,w)_j = p_j\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i, alt_j \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i, p_j = alt_j)\mathbb{P}(\{w : F_I(p_I,w)_{i,j} = (alt_i, alt_j)\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i, alt_j \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I)\mathbb{P}(alt_i)\mathbb{P}(alt_j)\mathbb{P}(\{w : F_I(p_I,w)_{i,j} = (alt_i, alt_j)\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I)\frac{1}{|\mathcal{A}|}\frac{1}{|\mathcal{A}|} \sum_{alt_i, alt_j \in \mathcal{A}} \mathbb{P}(\{w : F_I(p_I,w)_{i,j} = (alt_i, alt_j)\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I)\frac{1}{|\mathcal{A}|}\frac{1}{|\mathcal{A}|} \cdot 1 \\
&= \frac{1}{|\mathcal{A}|}\frac{1}{|\mathcal{A}|} \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I) \\
&= \frac{1}{|\mathcal{A}|}\frac{1}{|\mathcal{A}|} \cdot 1 = \frac{1}{|\mathcal{A}|^2}
\end{aligned}$$

Por el otro lado tenemos que:

$$\begin{aligned}
\mathbb{P}(B_i) &= \mathbb{P}(\{(p,w) : F_I(P_I,w)_i = p_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i\})\mathbb{P}(\{w : F_I(p_I,w)_i = alt_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I)\mathbb{P}(alt_i)\mathbb{P}(\{w : F_I(p_I,w)_i = alt_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I)\frac{1}{|\mathcal{A}|} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(\{w : F_I(p_I,w)_i = alt_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I)\frac{1}{|\mathcal{A}|} \cdot 1 \\
&= \frac{1}{|\mathcal{A}|} \sum_{p_I \in \mathcal{A}^I} \mathbb{P}(P_I(p) = p_I) \\
&= \frac{1}{|\mathcal{A}|} \cdot 1 = \frac{1}{|\mathcal{A}|}
\end{aligned}$$

Análogamente, tenemos que $\mathbb{P}(B_j) = \frac{1}{|\mathcal{A}|}$. Así:

$$\mathbb{P}(B_i \cap B_j)\mathbb{P}(B_i) \cdot \mathbb{P}(B_j)$$

Por lo que ambos conjuntos son independientes y con ello las dos funciones iniciales también son independientes. Como las funciones son independientes, y cada función es una indicatriz, entonces la ganancia es una suma de variables Bernoulli, con cada indicatriz una variable Bernoulli de parámetro $\frac{1}{|\mathcal{A}|}$. Así:

$$1_{\{F_I(P_I(p),\omega)_i=p_i\}} \sim Bernoulli\left(\frac{1}{|\mathcal{A}|}\right) \implies \sum_{i\in I^c} 1_{\{F_I(P_I(p),\omega)_i=p_i\}} \sim Binom\left(|I^c|, \frac{1}{|\mathcal{A}|}\right)$$

Como la distribución de la ganancia no depende de la estrategia, entonces todas las estrategias tienen igual distribución en cada I, y con ello, tienen igual ganancia. □

Corolario A.9 (Estrategia control) *Sea $\mathcal{A}^N$ un espacio de secuencias y $\mathcal{P}(\mathcal{A}^N)$ su σ-álgebra, entonces para cualquier probabilidad $\mathbb{P}$ la ganancia de la estrategia Random en $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P})$ es la misma.*

Demostración. Sea $\mathbb{P}$ una probabilidad, sea F la estrategia Random, y sea $\emptyset \subsetneq I \subsetneq \{1, ..., N\}$ un conjunto de índices, mostraremos primero que $1_{\{F_I(P_I(p),\omega)_i=p_i\}}$ y $1_{\{F_I(P_I(p),\omega)_j=p_j\}}$ son independientes. Igual que en el corolario anterior, esta es independencia es equivalente a probar que los conjuntos $B_i := \{F_I(P_I(p),\omega)_i = p_i\}$ y $B_j := \{F_I(P_I(p),\omega)_j = p_j\}$ son independientes. Tenemos que:

$$\begin{aligned}
\mathbb{P}(B_i \cap B_j) &= \mathbb{P}(\{(p,w) : F_I(P_I,w)_i = p_i\} \cap \{(p,w) : F_I(P_I,w)_j = p_j\}) \\
&= \sum_{p_I\in\mathcal{A}^I} \sum_{alt_i,alt_j\in\mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i, p_j = alt_j) \\
&\quad \cdot \mathbb{P}(\{(p,w) : F_I(p_I,w)_{i,j} = (alt_i, alt_j)\} | P_I(p) = p_I, p_i = alt_i, p_j = alt_j) \\
&= \sum_{p_I\in\mathcal{A}^I} \sum_{alt_i,alt_j\in\mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i, p_j = alt_j) \\
&\quad \cdot \mathbb{P}(\{w : F_I(p_I,w)_{i,j} = (alt_i, alt_j)\}) \\
&= \sum_{p_I\in\mathcal{A}^I} \sum_{alt_i,alt_j\in\mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i, p_j = alt_j) \frac{1}{|\mathcal{A}|}\frac{1}{|\mathcal{A}|} \\
&= \frac{1}{|\mathcal{A}|^2} \sum_{p_I\in\mathcal{A}^I} \sum_{alt_i,alt_j\in\mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i, p_j = alt_j) \\
&= \frac{1}{|\mathcal{A}|^2} \cdot 1 = \frac{1}{|\mathcal{A}|^2}
\end{aligned}$$

Y para las probabilidades individuales que:

$$\begin{aligned}
\mathbb{P}(B_i) &= \mathbb{P}(\{(p,w) : F_I(P_I,w)_i = p_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i)\mathbb{P}(\{(p,w) : F_I(p_I,w)_i = alt_i\} | P_I(p) = p_I, p_i = alt_i) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i)\mathbb{P}(\{w : F_I(p_I,w)_i = alt_i\}) \\
&= \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i)\frac{1}{|\mathcal{A}|} \\
&= \frac{1}{|\mathcal{A}|} \sum_{p_I \in \mathcal{A}^I} \sum_{alt_i \in \mathcal{A}} \mathbb{P}(P_I(p) = p_I, p_i = alt_i) \\
&= \frac{1}{|\mathcal{A}|} \cdot 1 = \frac{1}{|\mathcal{A}|}
\end{aligned}$$

Análogamente, para B_j tenemos que $\mathbb{P}(B_j) = \frac{1}{|\mathcal{A}|}$. Razonando como en el corolario anterior, tenemos que como las funciones son independientes, y cada función es una indicatriz, entonces la ganancia es una suma de variables Bernoulli, con cada indicatriz una variable Bernoulli de parámetro $\frac{1}{|\mathcal{A}|}$. Así:

$$1_{\{F_I(P_I(p),\omega)_i = p_i\}} \sim Bernoulli\left(\frac{1}{|\mathcal{A}|}\right) \implies \sum_{i \in I^c} 1_{\{F_I(P_I(p),\omega)_i = p_i\}} \sim Binom\left(|I^c|, \frac{1}{|\mathcal{A}|}\right)$$

Como la distribución de la ganancia no depende de la probabilidad, es decir, del modelo, entonces la estrategia Random tiene igual distribución de ganancia en todos los modelos, y con ello, igual ganancia en todos los modelos. □

Teniendo definiciones para estrategia y ganancia promedio, enunciamos la existencia de una extrategia de ganancia maximal:

Teorema A.10 (Extrategia de ganancia maximal) *Sea $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P})$ un espacio de probabilidad, entonces existe una estrategia determinista con mayor o igual ganancia promedio que cualquier otra estrategia en este espacio.*

Demostración. Sea F una estrategia cualquiera, y sea $\emptyset \subsetneq I_{conoc} \subsetneq \{1, ..., N\}$ un conjunto de indices, entonces podemos escribir la ganancia promedio como:

$$
\begin{aligned}
G_F(I) &= \sum_{p\in\mathcal{A}^N} \mathbb{P}(p)\mathbb{E}_\omega(\sum_{i\in I^c} 1_{\{w:F_I(P_I(p),\omega)_i=p_i\}}) \\
&= \sum_{i\in I^c}\sum_{p\in\mathcal{A}^N} \mathbb{P}(p)\mathbb{P}(\{w : F_I(P_I(p),\omega)_i = p_i\}) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I}\sum_{alt\in\mathcal{A}}\sum_{p:P_I(p)=p_I\wedge p_i=alt} \mathbb{P}(p)\mathbb{P}(\{w : F_I(P_I(p),\omega)_i = p_i\}) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I}\sum_{alt\in\mathcal{A}} \mathbb{P}(\{w : F_I(p_I,\omega)_i = alt\}) \sum_{p:P_I(p)=p_I\wedge p_i=alt} \mathbb{P}(p) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I}\sum_{alt\in\mathcal{A}} \mathbb{P}(\{w : F_I(p_I,\omega)_i = alt\})\mathbb{P}(P_I(p)=p_I, p_i = alt) \\
&\leq \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I} \max_{alt\in\mathcal{A}} \mathbb{P}(P_I(p)=p_I, p_i = alt) \sum_{alt\in\mathcal{A}} \mathbb{P}(\{w : F_I(p_I,\omega)_i = alt\}) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I} \max_{alt\in\mathcal{A}} \mathbb{P}(P_I(p)=p_I, p_i = alt)\cdot 1 \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I} \mathbb{P}(P_I(p)=p_I) \max_{alt\in\mathcal{A}} \mathbb{P}(p_i = alt|P_I(p)=p_I)
\end{aligned}
$$

Notando que la última cota no depende de F, mostraremos que existe una estrategia determinista F^* que tiene ganancia promedio esta última cota, y con ello, tiene o igual ganancia promedio que cualquier otra estrategia. Definamos la función g como:

$$
g : \mathcal{A}^I \times \Omega \to \mathcal{A}^I; \quad g(p_I,\omega)_i = (\arg\max_{alt\in\mathbb{A}} \mathbb{P}(p_i = alt|P_I(p)=p_I))[0]
$$

Usamos [0] al final del conjunto argmax para elegir el primer elemento según algún orden en $\mathcal{A}$, con ello, hacer que la elección sea determinista, pero cabe destacar que el siguiente cálculo no depende de que elemento del argmax usemos para la definición de g, sólo depende de que el elemente esté en el argmax. Definiendo $F^*_I = g$ tenemos que:

$$
\begin{aligned}
G_{F^*}(I) &= \sum_{p\in\mathcal{A}^N} \mathbb{P}(p)\mathbb{E}_\omega(\sum_{i\in I^c} 1_{\{w:g(P_I(p),\omega)_i=p_i\}}) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I}\sum_{alt\in\mathcal{A}} \mathbb{P}(\{w : g(p_I,\omega)_i = alt\})\mathbb{P}(P_I(p)=p_I, p_i = alt) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I}\sum_{alt\in\mathcal{A}} \mathbb{P}(\{w : g(p_I,\omega)_i = alt\})\mathbb{P}(p_i = alt|P_I(p)=p_I)\mathbb{P}(P_I(p)=p_I) \\
&= \sum_{i\in I^c}\sum_{p_I\in\mathcal{A}^I} \max_{alt\in\mathcal{A}} \mathbb{P}(p_i = alt|P_I(p)=p_I)\mathbb{P}(P_I(p)=p_I)
\end{aligned}
$$

Notando que la estrategia F^* definida anteriormente es determinista, se concluye la demostración. □

Corolario A.11 (Extrategia de ganancia minimal) *Sea* $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P})$ *un espacio de probabilidad, entonces existe una estrategia determinista con mayor o igual ganancia promedio que cualquier otra estrategia en este espacio.*

Demostración. Haciendo un razonamiento análogo a la demostración de la estrategia maximal, pero usando el argmin en vez del argmax, se obtiene el resultado. □

Corolario A.12 (Underdog es una estrategia maximal) *Sea* $(\mathcal{A}^N, \mathcal{P}(\mathcal{A}^N), \mathbb{P})$ *el espacio de probabilidad dado por el modelo Balanceado Exacto, i.e., la probabilidad uniforme entre el conjunto de las secuencias balanceadas exactamente, entonces la estrategia Underdog es una estrategia de ganancia maximal en este espacio.*

Demostración. Sea $\emptyset \subsetneq I \subsetneq \{1, ..., N\}$ un conjunto de índices, y F la estrategia Underdog, mostraremos que

$$F_I(p_I, \omega) = (\arg\max_{alt \in \mathbb{A}} \mathbb{P}(p_i = alt | P_I(p) = p_I))[k]$$

para algún k índice del conjunto argmax, y como la ganancia es independiente del elemento del argmax que se elija, entonces Underdog será una estrategia de gananacia maximal.

Primero notamos que si condicionamos en $P_I(p) = p_I$, está probabilidad será una uniforme entre las secuencias p tales que $P_I(p) = p_I$. Si llamamos N_{alt} a la función que cuenta los elementos alt de la secuencia como en el Corolario A.5, entonces tenemos que las secuencias p tales que $P_I(p) = p_I$ son aquellas tales que:

$$p \in \mathcal{A}^N : N_{alt}(P_{I^c}(p)) + N_{alt}(P_I(p)) = \frac{N}{|\mathcal{A}|}, \forall alt \in \mathcal{A}$$

es decir, las secuencias que tengan una cierta cantidad de elementos alt en I^c. Como esta condición no depende del orden de los elementos en I^c y sólo depende de la cantidad de cada elemento, entonces la probabilidad de que en el indice $i \in I^c$ esté el elemento alt es:

$$N_{alt}(P_{I^c}(p)) = \frac{\frac{N}{|\mathcal{A}|} - N_{alt}(P_I(p))}{|I^c|}$$

y esta probabilidad se maximiza cuando el término $N_{alt}(P_I(p))$ es mínimo, es decir, en el (o los) elementos que tienen la mínima frecuencia en $P_I(p)$. Notando que $P_I(p)$ son las respuestas por conocimiento en el caso de la estrategia Underdog, tenemos que la elección de Underdog está en el argmax definido para la estrategia maximal, por lo que la estrategia Underdog tiene la misma ganancia que la estrategia maximal determinista, y con ello, también es una estrategia de ganancia maximal. □

Anexo B

Cálculo del número de simulaciones

Sea $(\mathcal{A}^N, \mathcal{P}(A), \mathbb{P})$ un modelo de contrucción y F una estrategia en $\mathcal{A}^N$. Sea también N_{nivel} un número de ítems que se deben resolver por conocimiento dados por un nivel, entonces simularemos la función de ganancia para este nivel:

$$G_{F,nivel} : \{I \in \mathcal{P}(\mathcal{A}^N) : |I| = N_{nivel}\} \times \mathcal{A}^N \times \Omega \to \mathbb{N}; \quad (I, p, \omega) \longmapsto G_F(I)(p, \omega)$$

Como $G_{F,nivel}$ es una v.a. acotada (la lantidad de puntos de ganancia está acotada por la cantidad de ítem por que se responden por estrategia), entonces $\sigma^2 = Var(G_{F,nivel}) < \infty$.

Usando el Teorema central de límite, tenemos que si $X_1, ..., X_n$ son n muestras independientes de $G_{F,nivel}$ entonces:

$$\frac{X_1 + ... + X_n}{n} \sim \mathbb{E}(G_{F,nivel}) + \frac{\sigma}{\sqrt{n}} Z$$

Con Z una normal estándar. Aproximaremos el promedio de la ganancia del nivel con un orden de una décima de punto, es decir, 0.1. Para esto, primero calcularemos el número de simulaciones necesarias para que la desviación estándar del error $\frac{\sigma}{\sqrt{n}}$ sea de este orden con una confianza de 95 %:

$$\frac{\sigma}{\sqrt{n}} = \frac{orden}{2} \implies n = ((2 * \sigma)/orden)^2$$

Se aproximó la desviación estándar con un número de 10.000 simulaciones, y luego calcularon las máximas desviaciones estándar de las ganancias de cada estudio (Estudio 1, 4 opciones: 6.23; Estudio 2, 4 opciones: 4.35; Estudio 3, 4 opciones: 4.88; Estudio 1, 3 opciones: 10.43; Estudio 2, 3 opciones: 4.89; Estudio 3, 3 opciones: 6.58; Estudio 1, 5 opciones: 4.26; Estudio 2, 5 opciones: 4.01; Estudio 3, 5 opciones: 4.24).

Entonces se calculó el número de simulaciones necesarias (Estudio 1, 4 opciones: 15532; Estudio 2, 4 opciones: 7569; Estudio 3, 4 opciones: 9526; Estudio 1, 3 opciones: 43589; Estudio

2, 3 opciones: 9583; Estudio 3, 3 opciones: 17348; Estudio 1, 5 opciones: 7291; Estudio 2, 5 opciones: 6459; Estudio 3, 5 opciones: 7199).

Y luego se simularon las ganancias usando los siguientes número de simulaciones: 10.000 (para el Estudio 2 en 4 opciones, el Estudio 3 en 4 opciones, el Estudio 2 en 3 opciones, el Estudio 1 en 5 opciones, el Estudio 2 en 5 opciones y Estudio 3 en 5 opciones), 20.000 (para el Estudio 1 en 4 opciones y el Estudio 3 en 3 opciones) y 45.000 (para el Estudio 1 en 3 opciones).