Universidad de Lima

Facultad de Psicología

Carrera de Psicología

UN EXPERIMENTO MONTE CARLO SOBRE EL EFECTO DE LA DISTANCIA ENTRE LAS MEDIAS ARITMÉTICAS DE LOS PARÁMETROS DE HABILIDAD DE DOS GRUPOS POBLACIONALES EN LA CONFIABILIDAD

Tesis para optar por el Título Profesional de Licenciado en Psicología

Ángel Christopher Zegarra López

Código 20133289

Asesor

Andrés Burga León

Lima – Perú

Febrero de 2020

UNIVERSIDAD DE LIMA
MCMLXII
SCIENTIA ET PRAXIS

UN EXPERIMENTO MONTE CARLO SOBRE EL EFECTO DE LA DISTANCIA ENTRE LAS MEDIAS ARITMÉTICAS DE LOS PARÁMETROS DE HABILIDAD DE DOS GRUPOS POBLACIONALES EN LA CONFIABILIDAD

RESUMEN

El objetivo de esta investigación es comparar la confiabilidad obtenida a partir de la aplicación de un test on target para toda una población con aquella obtenida al utilizar tests on target para dos grupos específicos que componen dicha población, en situaciones que varían de acuerdo a la distancia entre las medias aritméticas de los parámetros de habilidad de ambos grupos. A través de simulaciones Monte Carlo, se diseñó un estudio experimental en donde las variables independientes fueron el target de la evaluación (un solo test enfocado en la población y dos tests enfocados en los grupos que la componen), y la distancia entre medias aritméticas de parámetros de habilidad (0.5, 1.0, 1.5 y 2.0 desviaciones estándar). La variable dependiente fue la confiabilidad, estimada a partir del índice de confiabilidad de separación de personas. Los datos fueron simulados sobre la base del modelo Rasch, considerando 5000 casos para cada escenario, con niveles de habilidad muestreados partir de una distribución normal estándar; e instrumentos de 40 ítems con parámetros de dificultad equidistantes, en el rango de -2.5 a 2.5 desviaciones estándar alrededor de la media aritmética de la distribución de habilidad. 1000 réplicas fueron establecidas para cada uno de los 8 escenarios formados por la intersección entre las variables independientes. Los resultados fueron contrastados a través de un análisis de varianza factorial y una prueba de comparaciones múltiples diseñada para los casos en donde no se cumple el supuesto de homoscedasticidad. Los hallazgos indican que existen diferencias estadísticamente significativas entre la confiabilidad obtenida a partir de tests enfocados en la población y en los grupos específicos, en cada uno de los escenarios simulados de distancia entre medias aritméticas; sin embargo, el tamaño del efecto denota que dicha diferencia es ínfima e irrelevante en la práctica.

Palabras clave: *Confiabilidad, Targeting, Modelo Rasch, Diseño de instrumentos*.

ABSTRACT

The aim of this study is to compare the reliability estimated through the application of a single test targeted on a whole population and two tests targeted on two specific groups that compose the population, in situations that vary according to the difference between the ability parameters means from each group. A Monte Carlo experiment was carried out considering the target of the assessment (targeting the population and targeting two specific groups) and the difference between ability parameters arithmetic means (0.5, 1.0, 1.5 and 2.0 standard deviations) as independent variables. Reliability was set as the dependent variable, estimated through the person separation reliability index. The data was generated according to the dichotomous Rasch model. A sample of 5000 subjects were simulated and their true ability parameters where sampled from a standard normal distribution. A test length of 40 items was set for every scenario and their difficulty parameters were simulated as an equidistant sequence from -2.5 to 2.5 standard deviations around the ability parameters arithmetic means. 1000 replicas were executed for each of the 8 experiment conditions established by the intersection between independent variables. Results were analyzed using a robust version of the factorial analysis of variance and multiple comparisons tests designed specifically for situations where the homogeneity of variance assumption is not met. Findings show that there is a statistically significant difference between the reliability of the measures estimated through a population targeted test and through group targeted tests, in every scenario of distance between ability parameters means. Nevertheless, the effect size shows that this difference is not relevant on practice.

Key words: *Reliability*, *Targeting*, *Rasch model*, *Test design.*

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE APÉNDICES

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

El desarrollo de instrumentos de medición ha permitido el avance de diversas disciplinas científicas. La astronomía, por ejemplo, atravesó un desarrollo considerable entre los siglos XVII y XVIII gracias a la invención del telescopio. Del mismo modo, muchos de los avances en neurociencias sucedieron gracias a la invención de tecnologías como la tomografía por emisión de positrones y la imagen por resonancia magnética funcional (Urbina, 2014). En psicometría, el equivalente a estos instrumentos es el test psicométrico, el cual posibilita la medición indirecta de constructos psicológicos no observables a través de una muestra de comportamientos asociados a un dominio específico (Cohen y Swerdlik, 2010). De esta manera, es posible estudiar fenómenos psicológicos y proponer modelos teóricos que permitan describir, explicar, predecir y modificar la conducta (Ciccarelli y White, 2018).

No obstante, es importante reconocer que todo proceso de medición se encuentra expuesto a cierto grado de error y fluctuación. Incluso los instrumentos más refinados de las ciencias físicas son propensos a estos errores; por ello, dichos campos emplean métodos apropiados de calibración para garantizar la precisión de sus medidas (Urbina, 2014). En el caso de las ciencias del comportamiento, la medición a través del uso de tests psicométricos es más propensa a errores debido a la naturaleza de los constructos psicológicos latentes y a distintos factores que pueden intervenir durante el proceso de evaluación (Coulacoglou y Saklofske, 2017). Por esta razón, antes de realizar inferencias sobre la base de las respuestas a un test es indispensable asegurar que estas exhiban cierto grado de consistencia y precisión (Geisinger, 2013). A esta propiedad de las respuestas se le denomina *confiabilidad* (Cooper, 2019).

1.1 Descripción del problema

La confiabilidad (o también denominada *precisión*) es un principio psicométrico que se refiere a la consistencia de puntuaciones obtenidas a través de múltiples réplicas del proceso de evaluación (American

Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 2014). La necesidad de evaluar dicho grado de consistencia surge porque la medición de atributos psicológicos a través de instrumentos psicométricos es imperfecta, pues este proceso es susceptible a distintas fuentes de error (Brennan, 2001; Haertel, 2006; Kaplan y Saccuzzo, 2017; Urbina, 2014).

La presencia de errores de medición compromete la validez de las inferencias realizadas sobre la base de las puntuaciones de un test (Howitt y Cramer, 2017; Urbina, 2014). Esto se debe a que estas inferencias dependen del supuesto de que los individuos exhiben cierto grado de consistencia en sus respuestas a través de administraciones independientes del proceso de evaluación (AERA et al., 2014). Por este motivo, la confiabilidad es reconocida como un requisito indispensable para la validez (Cooper, 2019; Geisinger, 2013).

En este sentido, la problemática asociada a la imprecisión e inconsistencia de puntuaciones involucra a cualquier disciplina que utilice el test psicométrico como fuente principal de información en ámbitos académicos y aplicados (Barker, Pistrang y Elliot, 2016; Coaley, 2010; Howitt y Cramer, 2017; Urbina, 2014). Efectivamente, los resultados, interpretaciones y conclusiones de la investigación científica psicológica se encontrarán sesgados si carecen de mediciones precisas sobre los atributos estudiados (Brough y Hawkes, 2019; Coolican y Kelly, 2014; Finch y French, 2019). Del mismo modo, la imprecisión de los resultados de las evaluaciones psicométricas comprometerá todo proceso de toma de decisiones que se realice en ámbitos aplicados de las ciencias del comportamiento (AERA et al., 2014; Cooper, 2019).

Más aún, cuando un proceso de toma de decisiones sobre la base de las puntuaciones de un test implica una serie de consecuencias de alto impacto, la necesidad de contar con un mayor grado de confiabilidad incrementa (Kaplan y Saccuzzo, 2018; Howitt y Cramer, 2017). Este es el caso de las evaluaciones educativas a gran escala, cuyos resultados son utilizados para

el monitoreo, evaluación y rendición de cuentas del sistema educativo, así como para la reforma curricular, la promoción de la equidad y la mejora en los procesos de enseñanza y aprendizaje a nivel mundial (Lietz y Tobin, 2016; Tobin, Nugroho y Lietz, 2016; United Nations Educational, Scientific and Cultural Organization [UNESCO], 2018; 2019).

En general, las evaluaciones educativas a gran escala son programas estandarizados de evaluación en donde participa una extensa cantidad de estudiantes con distintas características socioculturales (Almond, Lehr, Thurlow y Quenemoen, 2002). Sin embargo, cuando estas características son muy divergentes, pueden representan una potencial fuente de error de medición (Kubiszyn y Borich, 2013; The British Psychological Society [BPS], 2017). Efectivamente, la AERA et al. (2014) afirman que si existen grupos con características muy distintas a las del resto de la población, es posible que el grado de confiabilidad que se obtiene al considerar a toda la población difiera del obtenido al considerar únicamente aquel grupo particular.

Una explicación a este fenómeno se encuentra en el propio diseño de las evaluaciones educativas a gran escala. Para comprender esta idea, es necesario entender que un proceso de medición implica obtener información particular sobre un grupo de personas, al cual se denomina como *target* (Wright y Douglas, 1975). La calidad de la medición dependerá del grado en que el instrumento empleado para la evaluación brinde información relevante acerca de dicho target (Wright y Stone, 1979).

En las evaluaciones educativas a gran escala, el target suele delimitarse como un grupo de individuos ubicados en un nivel relativo de aprendizaje (UNESCO, 2019). De la misma manera, el contenido de los instrumentos de evaluación se desarrolla sobre la base de un plan curricular que delimita los conocimientos y habilidades que debería dominar el target dado el nivel en el que se encuentra (Sireci y Gándara, 2016). En otras palabras, los ítems utilizados en estas evaluaciones cubren el rango de habilidad esperado de la población objetivo en función a estándares curriculares establecidos (Mendelovits, 2017).

El problema es que el target designado en las evaluaciones educativas a gran escala involucra a un extenso número de estudiantes de diferentes contextos socioculturales y con diferentes niveles de dominio sobre el contenido evaluado (Cresswell, Schwantner y Waters, 2015). En este escenario, el instrumento puede ser considerado apropiado para algunos grupos de la población, pero no para aquellos cuyo nivel de dominio del contenido difiere en gran medida del nivel para el cual fue diseñado el instrumento.

En el contexto del modelo de medición Rasch, uno de los motivos de esta situación se atribuye a la presencia de ítems *off target* en el instrumento de evaluación. Dicha denominación se otorga a un ítem cuyo nivel de dificultad se aleja del rango de habilidad de las personas, es decir, o es muy difícil o muy fácil para la mayoría de evaluados (Bond y Fox, 2015). Estos ítems comprometen la confiabilidad de las puntuaciones, pues no brindan información relevante acerca del nivel de habilidad en el que se encuentran los individuos, incrementan el error de estimación de parámetros y evocan conductas como la adivinación, estilos de respuesta o falta de interés en la evaluación (Bond y Fox, 2015; Ingebo, 1997; Wright y Stone, 1979).

En reconocimiento de esta problemática, algunos programas de evaluación educativa estandarizada comenzaron a implementar Tests Adaptativos Informatizados [TAI] (Kirsch, Lennon, von Davier, Gonzales y Yamamoto, 2013; Luecht, 2014). Los TAI son evaluaciones computarizadas cuya característica principal es el empleo de un algoritmo que asigna secuencialmente ítems con un determinado nivel de dificultad apropiado para el nivel de habilidad del evaluado (Beiser, Vu y Gibbons, 2016; Breithaupt y Hare, 2016; Loe, Stillwell y Gibbons, 2017; Urbina, 2014). Como cada persona se enfrenta únicamente a los ítems que brindan mayor información sobre su nivel de habilidad, esta práctica resulta en un incremento considerable de la precisión de las mediciones a comparación de los métodos tradicionales (Linacre, 2000; Luecht, 2014).

No obstante, el uso de metodologías de evaluación computarizada también presenta desventajas en distintos aspectos, especialmente en el contexto de

las evaluaciones educativas a gran escala (Walker, 2017). Por ejemplo, al evaluar a un extenso grupo de individuos en diversos contextos socioculturales, el grado de familiaridad que cada uno presenta en relación a la tecnología designada para la evaluación puede implicar una potencial fuente de error de medición (AERA et al., 2014; Luecht, 2014; Walker, 2017). Además, la implementación de una evaluación computarizada a gran escala requiere considerar una serie de aspectos logísticos como el perfil de profesionales requeridos para la aplicación, la anticipación ante posibles fallas de hardware y software, y la disponibilidad de recursos tecnológicos en los contextos en donde se desea evaluar (Breithaupt y Hare, 2016; Cohen y Swerdlik, 2009).

Una alternativa análoga a los TAI en el formato tradicional de lápiz y papel consiste en desarrollar instrumentos de evaluación a través de la estrategia *targeting* (Berezner y Adams, 2017). Dicha estrategia consiste en el diseño de pruebas constituidas por ítems que cubren el rango de habilidad de un grupo de personas, también denominadas pruebas *on target* (Boone, Staver y Yale, 2014; Cavanagh y Waugh, 2011; Wright y Stone, 1979). De esta manera, los ítems utilizados maximizan la información acerca del nivel de habilidad de las personas y minimizan el error de estimación (Bond y Fox, 2015; Luecht, 2014).

Dados los postulados de diversos autores en el marco del modelo de medición Rasch (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Ingebo, 1997; Linacre, 2000; Wright y Stone, 1979; 1999), utilizar tests *on target* para cada grupo relevante de la población debería proporcionar puntuaciones con un mayor grado de confiabilidad que el obtenido a través de un test único diseñado para cubrir el rango de habilidad de toda la población. Además, como la diferencia entre el nivel de habilidad de los grupos implica una fuente importante de error de medición, conforme dicha diferencia incremente, la confiabilidad obtenida al utilizar tests on target para cada grupo debería ser mayor.

Desafortunadamente, en la literatura especializada no existen estudios que contrasten esta hipótesis a través de estrategias empíricas. Además, el diseño requerido para contrastar dicha hipótesis es sumamente complejo, pues implica contemplar el control de distintas variables que pueden influir en la estimación de la confiabilidad como el tamaño de muestra, longitud de la escala, variabilidad de los datos, entre otros (Linacre, 2019a). Incluso, sería necesario un control sobre la variable de interés, la diferencia entre los niveles de habilidad de los grupos que componen la población, lo cual en la práctica es imposible, pues las aproximaciones contemporáneas para determinar la medida de habilidad de una persona solo permiten *estimar* la habilidad latente verdadera.

En otras palabras, las estrategias de recolección de datos empíricos a través del uso de instrumentos psicométricos aplicados directamente a personas naturales no permitirían recrear los escenarios necesarios para contrastar la hipótesis. Por estos motivos, establecer un estudio que permita abordar esta temática requiere de una estrategia que permita ejercer un alto grado de control sobre las variables de interés y las variables extrañas que podrían sesgar los resultados. Afortunadamente, existe una estrategia ampliamente utilizada en diversos campos científicos y, particularmente, en la psicometría, para abordar estas situaciones, los experimentos de simulaciones Monte Carlo (Feinberg y Rubright, 2016).

Fundamentalmente, un experimento de simulaciones busca resolver preguntas complejas a partir de la generación de datos aleatorios que pretenden representar ciertas condiciones de estudio ideales (Harwell, Stone, Hsu y Kirisci, 1996). De esta manera, variables como los verdaderos niveles de habilidad latente de cada grupo y su distribución pueden ser fácilmente controladas. En este sentido, si se establece que los niveles de habilidad de cada grupo se distribuyen normalmente, como varios fenómenos en la realidad (Grami, 2020), entonces la diferencia entre ambos puede definirse como la distancia entre los parámetros de *localización* de dichas distribuciones, los cuales son representados por sus respectivas medias aritméticas (Grami, 2020).

En síntesis, el presente estudio utiliza el método de simulaciones Monte Carlo para establecer distintos escenarios que permitan determinar si utilizar tests on target para grupos específicos de la población implicará un incremento en la confiabilidad de las puntuaciones, tal y como se establece en el marco del modelo de medición Rasch (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Ingebo, 1997; Linacre, 2000; Wright y Stone, 1979; 1999). Al ser una primera aproximación a esta temática, se pretender evaluar la situación particular de una población compuesta únicamente por dos grupos relevantes. De este modo, se pretende responder a la pregunta: ¿cuánta distancia entre las medias aritméticas de los parámetros de habilidad de dos grupos poblacionales es necesaria para observar una ganancia significativa en la confiabilidad obtenida al emplear tests on target para cada grupo poblacional en comparación con la confiabilidad obtenida al utilizar una sola prueba on target para toda la población, en el contexto del modelo de medición Rasch?

1.2 Justificación y relevancia

Al delimitar un problema de investigación, es indispensable justificar la importancia del estudio en función a las implicancias teóricas y prácticas de sus posibles resultados (Brough y Hawkes, 2019; Roni, Merga y Morris, 2020). En congruencia con esta idea, la relevancia de determinar si el uso de tests on target para grupos poblacionales resulta en un incremento significativo en la confiabilidad de las puntuaciones se sustenta en un aporte teórico para la psicometría, y en un aporte práctico para los procesos de diseño y construcción de instrumentos en el contexto de las evaluaciones educativas a gran escala.

Las implicancias teóricas del estudio se derivan de la propia naturaleza de la investigación científica. Un marco teórico integra una conceptualización concisa acerca de un fenómeno en particular, a partir del cual es posible

plantear una serie de preguntas de investigación cuya resolución implica un aporte al respectivo campo de estudio (Bordens y Abbott, 2018; Marczyk, DeMatteo y Festinger, 2005). En este sentido, responder a la pregunta de investigación no solo aportará a la ampliación de dicho marco teórico, sino que posibilitará la implementación de otras preguntas e hipótesis sobre el fenómeno de estudio (Salkind, 2018).

En el marco del modelo de medición Rasch, diversos autores afirman que la estrategia targeting para el diseño de instrumentos en función al nivel de habilidad de las personas incrementa el grado de confiabilidad de las puntuaciones (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Íngebo, 1997; Linacre, 2000; Wright y Stone, 1979; 1999). Con el objetivo de aportar al respectivo marco teórico, en este estudio se pretende contrastar la hipótesis descrita a través de un experimento Monte Carlo que permita determinar la eficacia de la estrategia de elaborar pruebas on target para dos grupos que componen una población, al comparar la confiabilidad obtenida en esta evaluación con aquella obtenida al diseñar un único test on target para toda la población.

Las implicancias prácticas de la investigación se describen en relación a cómo los resultados del estudio serán de utilidad en la resolución de problemas concretos en ámbitos aplicados (Bordens y Abbott, 2018; Brough y Hawkes, 2019). El problema que se aborda en esta investigación surge en el contexto de las evaluaciones educativas a gran escala. Al evaluar a una población extensa compuesta por grupos con características muy distintas al resto, es posible que el grado de confiabilidad obtenido al evaluar a toda la población difiera del obtenido al considerar únicamente las puntuaciones de los grupos particulares (Coulacoglou y Saklofske, 2017).

En otras palabras, cabe la posibilidad de que al evaluar la confiabilidad de las puntuaciones de toda la población se obtenga un grado adecuado, pero al analizar particularmente las puntuaciones de los grupos relevantes se obtenga un grado de confiabilidad inapropiado para alguno de ellos (AERA et al., 2014). Un bajo grado de confiabilidad compromete la validez de las

inferencias de los resultados de una evaluación y sesga todo proceso de toma de decisiones sobre la base de las puntuaciones de un test (Geisinger, 2013). Incluso, esta problemática es más grave cuando los resultados de una evaluación implican consecuencias importantes para los evaluados (AERA et al., 2014; BPS, 2017).

En vista de que los resultados de las evaluaciones educativas a gran escala son utilizados para el establecimiento de reformas sociales, económicas, políticas y educativas (Klieme, 2013; Lietz y Tobin, 2016), es imprescindible garantizar un alto grado de confiabilidad para cada grupo relevante de la población (AERA et al., 2014; International Test Comission [ITC], 2013a; 2013b; 2018; Educational Testing Service [ETS], 2014). Por esta razón, determinar si el uso de pruebas on target para cada grupo de la población representa una estrategia de diseño y construcción de instrumentos apropiada para el incremento de los niveles de confiabilidad.

CAPÍTULO II: MARCO TEÓRICO

El conocimiento científico se desarrolla sobre la base de un marco coherente de evidencia y teoría (Kosso, 2011; Carey, 2011). En este sentido, la adquisición de nuevo conocimiento se constituye a partir de la observación de los fenómenos de estudio en relación a un marco teórico que permita describir y explicar su naturaleza (Bordens y Abbott, 2018; Kosso, 2011). En congruencia con este principio de la investigación científica, el presente estudio se establece sobre el marco del modelo de medición Rasch, una propuesta que pretende establecer altos estándares para la medición en las ciencias sociales (Bond y Fox, 2015).

Para comprender a profundidad los postulados del modelo Rasch es necesario conocer acerca del contexto histórico y metodológico en el cual surge. Por este motivo, en la primera sección de este capítulo se presenta una breve introducción a la discusión epistemológica sobre la medición en psicología y el rol que ocupa el modelo Rasch en ella. En la segunda sección, se describen las características principales del modelo Rasch para ítems dicotómicos. En la tercera sección se introduce el concepto de confiabilidad, uno de los principios fundamentales de la medición psicológica. En la cuarta sección, se describe la aproximación del modelo Rasch con respecto a la confiabilidad, introduciendo los métodos de estimación de dicha propiedad y el concepto de targeting. Finalmente, en la última sección se realiza una breve introducción a la lógico de los experimentos Monte Carlo en psicometría.

2.1 La medición en psicología y el modelo Rasch

La familia de modelos Rasch se origina en el trabajo de George Rasch, un matemático danés que en 1960 propuso un modelo de medición para respuestas a ítems dicotómicos en la obra *Probabilistic Models for Some Intelligence and Attainment Tests* (Andrich y Marais, 2019; Christensen, Kreiner y Mesbah, 2013; Paek y Cole, 2020). Benjamin Wright (Rasch, 1960/1980) considera a esta publicación como el trabajo más importante en

psicometría desde los artículos de Thurstone entre los años 1925 y 1929. Aunque los modelos presentados en esta obra se orientan a la evaluación de la inteligencia y algunos aspectos de lectura, Wright (Rasch, 1960/1980) afirma que los métodos presentados en este trabajo trascienden la medición psicológica y educativa, pues representan los principios en los que se fundamenta la objetividad, reproducibilidad y el conocimiento científico, aquellos principios esenciales de la medición en sí misma.

Prima facie, la medición es un concepto importante en las ciencias físicas y considerada como un método privilegiado para la adquisición de información sobre el mundo (Mari y Wilson, 2014; Maul, Torres-Irribarra y Wilson, 2016). Sin embargo, la evolución del entendimiento sobre la naturaleza de la medición ha sido influenciada por factores históricos y prácticos, resultando en la coexistencia de varias definiciones sobre este concepto (Humpry, 2013; Mari, Maul, Torres-Irribarra y Wilson, 2017). Mari et al. (2017) presentan algunos ejemplos de este suceso:

> "La medición es cualquier método por el cual una correspondencia única y recíproca se establece entre todas o algunas las magnitudes de un tipo y todos o algunos de los números, integrales, racionales o reales, dado el caso" (Russell, 1903/2010, p. 176)
>
> "La medición es el proceso de asignar números que representan cualidades" (Campbell, 1920, p. 267)
>
> "La medición es la correlación entre números de entidades que no son números" (Nagel y Hempel, 1931, p. 313)
>
> "La medición es la asignación de numerales a objetos o eventos de acuerdo a ciertas reglas" (Stevens, 1946, p. 677)
>
> "La medición es el descubrimiento o estimación de relaciones numéricas (ratios) entre magnitudes de un atributo cuantitativo y una unidad" (Michell, 1999, p. 76).

El estudio sobre la naturaleza de la medición sigue siendo materia de debate en la literatura contemporánea (e.g., Mari, Maul, Torres-Irribarra y Wilson, 2016). Lamentablemente, en las ciencias del comportamiento muy pocos

especialistas reconocen la importancia del concepto de medición para el estudio de los fenómenos psicológicos (Bond y Fox, 2015; Borsboom, 2005; Michell, 2001). En contraste, el modelo Rasch se encuentra estrictamente relacionado con la noción de medición en psicología; sin embargo, para comprender dicha relación, es importante analizar los orígenes del paradigma predominante de la medición psicológica, la postura de Stevens (1946).

A finales del siglo XIX, existían tres perspectivas principales en el campo de la teoría de la medición (McGrane, 2015). En primer lugar, la tradición *sistemática*, representada por los trabajos de James Maxwell (1873), uno de los físicos más importantes de todos los tiempos, se basa en la proposición de dos factores que componen la medición: la unidad de medida, que representa una cantidad física como estándar de referencia, y el ratio que indica el número de veces que la unidad debe darse para obtener la cantidad requerida. En segundo lugar, la perspectiva *representacional*, liderada por Norman Campbell (1921), un filósofo de la ciencia que definió la medición como el proceso de descubrir relaciones entre propiedades empíricas y objetos de modo que puedan asignarse a estos una serie de números que los representen adecuadamente. En tercer lugar, la aproximación *operacional*, liderada por los trabajos del ganador del premio nobel en física, Percy Bridgman (1927), sostiene que la medición es una asociación operacional entre un número y una cantidad, de modo que todo concepto es sinónimo de un conjunto de operaciones realizadas para identificarlo. En años posteriores, esta aproximación fue reinterpretada como toda operación precisamente especificada que produce un número (Dingle, 1950).

En este contexto de incertidumbre sobre la naturaleza de la medición, Stevens (1946) propuso definir el concepto como la asignación de numerales a objetos o eventos en función a reglas determinadas. La perspectiva representacional influenció en gran medida a su proposición; sin embargo, Campbell (1921) defendía la medición como asignación de numerales siempre y cuando se represente apropiadamente la relación empírica entre propiedades físicas y matemáticas al satisfacer los criterios

de *aditividad*, como parte de los criterios lógicos de la medición presentados en la tabla 2.1.

Tabla 2.1

Los requerimientos lógicos para la asignación de numerales de Campbell (1921)

N	Criterio
I	Si $A > B$, entonces $B < A$
II	Si $A > B$ y $B > C$, entonces $A > C$
III	$A = B$ solo si $A \geq B$ y $A \leq B$
IV	Si $A = B$, se cumple que si $A > C$, entonces $B > C$
V	Si $A = B$, se cumple que si $A < C$, entonces $B < C$
VI	Si $A = B$ y $B = C$, entonces $A = C$
VII	Si $A = B$m entonces $B = A$
VIII	Si $A = A'$ y $B > 0$, entonces $A + B > A'$
IX	Si $A + B = X, entonces\ B + A = X$
X	Si $A = A'$ y $B = B'$, entonces $A + B = A' + B'$
XI	$(A + B) + C = A' + (B' + C')$
XII	Si $B = 0$, entonces $A + B = A'$

Nota. *A*' y *B*' representan los numerales asignados a *A* y *B*. Adaptado de "The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples," por T. W. Reese, 1943, *Psychological Monographs, 55*(3), pp. 6-20. Copyright 1945 por The American Psychological Association.

En contraste, Stevens (1946) no reconoció la aditividad como condición necesaria para la asignación de numerales. En su lugar, el autor propuso que existía la posibilidad de asignar numerales a objetos o eventos que representen algún tipo de relación empírica no aditiva. Como fundamento para esta postura, Stevens (1946) adoptó los postulados del operacionalismo, que sustentaban que cualquier regla de asignación de

numerales a objetos o eventos pueda considerarse como una representación numérica de, por lo menos, una relación de *equivalencia* definida operacionalmente por la propia regla (Michel, 1997).

Por ejemplo, si se clasificara a un grupo de personas por sexo, a las mujeres se les asignaría un numeral X' y a los hombres Y'. De esta manera, se obtendría una relación empírica no aditiva de equivalencia, expresada como "ser del mismo sexo", que ocurriría si a dos personas A y B se les asignara el mismo numeral X' o Y' (Michell, 1990). Además, como toda operación que deriva en un número es considerada una medición en el paradigma operacionalista (Dingle, 1950), la asignación misma de numerales representa una medición en sí misma que siempre derivará en, por lo menos, la relación de equivalencia descrita en el ejemplo (Michell, 1997).

En función a las posibles relaciones que se pueden establecer a partir de la asignación de numerales, Stevens (1946) propuso cuatro "escalas de medición" (presentadas en la tabla 2.2). La equivalencia numérica, considerada la relación más básica que siempre producirá una asignación de numerales, constituye la escala denominada *nominal*, en la cual un número indica la identidad o diferenciación respecto a un atributo o categoría. Cuando los numerales permiten establecer relaciones de orden entre categorías, estos números corresponden a una escala *ordinal*. Si los numerales asignados permiten representan una equidad en las diferencias entre objetos con respecto a un atributo, se considera una escapa de *intervalo*; además, es posible asignar una unidad de medida y un punto cero de manera arbitraria. Finalmente, aquellos numerales cuya asignación representa apropiadamente una relación aditiva (que cumple los criterios de Campbell, 1920), se denomina como una escala de $razón$; en este nivel, el valor cero indica la ausencia del atributo medido. Estas cuatro escalas, aunque no las únicas propuestas por el autor, definen las reglas de asignación, los análisis estadísticos y las inferencias apropiadas para los numerales (Michell, 1990).

Tabla 2.2

Las escalas de medición de Stevens (1946)

Escala	Operaciones empíricas básicas	Estructura matemática grupal	Estadísticos permitidos
Nominal	Determinación de equivalencia	Grupo de permutación $x' = f(x)$ $f(x)$ es cualquier substitución uno a uno.	Número de casos Moda Correlación de contingencia
Ordinal	Determinación de mayor o menor	Grupo isotónico $x' = f(x)$ $f(x)$ es cualquier función monotónica creciente	Mediana Percentiles
Intervalo	Determinación de equivalencia de intervalos o diferencias	Grupo lineal general $x' = ax + b$	Media Desviación estándar Correlación *rank-order* Correlación *product-moment*
Razón	Determinación de equivalencia de ratios	Grupo de similitud $x' = ax$	Coeficiente de variación

Nota. Todo nivel permite realizar las operaciones empíricas básicas y análisis estadísticos del nivel precedente. x' representa al numeral asignado para x. Adaptado de "On the theory of scales of measurement," por S. S. Stevens, 1946, *Science, 103*, p. 678. Copyright 1946 por The American Association for the Advancement of Science.

Unas décadas antes de la propuesta de Stevens (1946), James Maxwell y William Thomson, como representantes del British Association for the Advancement of Science, propusieron adoptar el sistema de unidades de medición denominado *centímetro-gramo-segundo* [CGS, por sus siglas en inglés], inicialmente propuesto por Carl Friedrich Gauss. Años posteriores, la propuesta sistemática fue discutida y académica hasta lo que actualmente se conoce como el *Sistema Internacional de Unidades* (para una descripción histórica detallada, revisar McGrane, 2015).

Mientras que la perspectiva sistemática se establecía como el modelo predominante en la medición física científica, la medición en psicología se

desarrolló sobre la base del modelo de Stevens (1946). La incongruencia entre ambas aproximaciones ha dado origen a una serie de discusiones y críticas en la literatura especializada (e.g., Humphry, 2011; Kyngdon, 2013; Michell, 1999; Sherry, 2011; Trendler, 2009); sin embargo, esta temática no suele ser difundida entre psicólogos y psicometristas (Bond y Fox, 2015; Borsboom, 2005; Michell, 2001).

Entre las diversas posturas en esta línea de investigación, Joel Michell (1999) considera que la psicología pretende adherirse a un marco científico con el cual es incongruente. En principio, dicha incongruencia ocurre porque las medidas de los atributos psicológicos se desarrollan sobre la base de un modelo distinto al de las ciencias físicas. Desde la perspectiva *realista* de Michell (2000), la medición en ciencias físicas consiste en el *descubrimiento o estimación de la razón de magnitud entre una cantidad y otra unidad de la misma cantidad, la cual es aditiva e invariante*.

Esta aproximación supone que los atributos implicados en un proceso de medición son de naturaleza estrictamente cuantitativa (Michell, 1999; 2009). La noción moderna de *cantidad* fue propuesta por Ludwig Hölder (1901) en *Die Axiome der Quantität und die Lehre vom Mass*, en donde reformuló la teoría de la medición de cantidades de Euclides y postuló una serie de axiomas que determinan las características de la estructura cuantitativa para atributos continuos (Michell y Ernst, 1996). En atributos discretos, la estructura cuantitativa implica tres condiciones de *ordinalidad* y seis condiciones de *aditividad* de la teoría clásica de la medición (Michell, 1990; Heene, 2013). Los axiomas para los atributos continuos y discretos se presentan en las tablas 2.3 y 2.4, respectivamente.

Tabla 2.3

Axiomas de cantidad de Hölder (1901)

N	Axioma
I	Dadas dos magnitudes, a y b, una y solo una de las siguientes afirmaciones es verdadera: a es idéntica a b $(a = b, b = a)$, a es mayor que b y b es menor que a $(a > b, b < a)$, o de manera inversa, b es mayor que a y a es menor que b $(b > a, a < b)$.
II	Para cada magnitud, existe una que es menor.
III	Para cada par ordenado de magnitudes (no necesariamente distintas), a y b, su suma, $a + b$, se encuentra bien definida.
IV	$a + b$ es mayor que a y mayor que b.
V	Si $a < b$, entonces existe una x y una y, de manera que $a + x = b$ e $y + a = b$.
VI	Siempre es verdadero que $(a + b) + c = a + (b + c)$.
VII	Cuando todas las magnitudes son divididas en dos clases de modo que cada magnitud pertenezca a una y solo una clase, ninguna clase se encuentre vacía y cualquier magnitud en la primera clase sea menor a cada magnitud que la segunda clase; entonces existe una magnitud ξ de manera que todo $\xi' < \xi$ se encuentre en la primera clase y todo $\xi'' > \xi$ pertenezca a la segunda clase.

Nota. Adaptado de "The axioms of quantity and the theory of measurement," por J. Michell y C. Ernst, 1996, *Journal of Mathematical Psychology, 40,* p. 238. Copyright 1996 por Academic Press.

Tabla 2.4

Las condiciones de ordinalidad y aditividad de las variables cuantitativas (positivas y discretas)

Propiedad	Condición	Descripción
Ordinalidad	Transitividad	Si $X \geq Y, Y \geq Z$, entonces $X \geq Z$
	Asimetría	Si $X \geq Y, Y \geq X$, entonces $X = Y$
	Conexión fuerte	$X \geq Y$ o $Y \geq Z$
Aditividad	Asociatividad	$X + (Y + Z) = (X + Y) + Z$
	Conmutatividad	$X + Y = Y + X$
	Monotonicidad	$X \geq Y$ si y solo si $X + Z \geq Y + Z$
	Solubilidad	Si $X > Y$, entonces existe un valor Z de modo que $X = Y + Z$
	Positividad	Si $X > Y$, entonces $X + Y > X$
	Condición arquimédica	Si $X > Y$, entonces existe un número natural n de modo que $nX \geq Y$ (en donde $1X = X$ y $(n + 1)X = nX + X$)

Nota. Cada una de estas condiciones son hipótesis demostrables (Heene, 2013). Adaptado de *An introduction to the logic of psychological measurement* (pp. 52-53), por J. Michell, 1990, New York, NY; Psychology Press. Copyright 2014 por Lawrence Erlbaum Associates.

Sobre la base de lo anterior, Michell (2000) detalla distintos argumentos para sustentar que la perspectiva de las ciencias físicas y la propuesta de Stevens (1946) son completamente incompatibles e, incluso, contradictorias. En primer lugar, la medición en las ciencias físicas se realiza sobre *atributos*; mientras que la postura de Stevens (1946) implica una asignación directa de numerales a objetos o eventos. Ante esto, Michell (2000) argumenta que la asignación de numerales se limita a simples lineamientos para realizar una *codificación* cuyo único objetivo es permitir la proposición inferencias sobre la base de medios numéricos.

En segundo lugar, el descubrimiento o estimación de las relaciones de magnitud entre atributos implica la existencia de estos en la naturaleza, esta es una característica fundamental para la perspectiva realista de la medición que defiende Michell (1999). Por el contrario, la asignación de numerales no justifica la existencia de los constructos psicológicos. Aunque el estatus ontológico de los atributos psicológicos es un tema ampliamente discutido en la literatura, este trasciende al foco del estudio (para una introducción breve, consultar Borsboom, Mellenbergh y van Heerden, 2003).

En tercer lugar, Michell (2009) afirma que, incluso si los atributos psicológicos existieran, no necesariamente tienen una naturaleza cuantitativa. En otros términos, no es posible afirmar que dichos atributos posean una estructura de relaciones que cumpla con los axiomas de cantidad. Para Michell (1999; 2008), los psicometristas se limitan a asumir, sin ningún sustento, que los atributos psicológicos existen y que poseen una estructura cuantitativa (McGrane, 2010; Osborne, 2010; Tendler, 2009). Esta práctica ha llevado al autor a clasificar a la psicometría como una *ciencia patológica*, aquella ciencia que avanza sobre la base de supuestos no demostrados empíricamente (Michell, 2000; 2008).

Sin lugar a dudas, la postura de Michell (1999) es extremadamente crítica; sin embargo, es importante considerar que los sistemas de medición en las ciencias físicas se consolidan sobre la base de un trabajo de siglos de refinamiento; mientras que, la psicología como disciplina científica apenas se originó a finales del siglo XIX (McGrane, 2010; Ciccarelli y White, 2018). De todas maneras, el consenso entre diversos autores en esta línea de investigación es que, si la psicología pretende alcanzar los mismos avances metodológicos que las ciencias físicas, debe abandonar el modelo de Stevens (1946) y adherirse a los principios de la teoría cuantitativa (McGrane; 2015; Tendler, 2009; Michell, 1999).

Afortunadamente, en 1964, Robert Duncan Luce y John Wilder Tukey propusieron una teoría congruente con la perspectiva de medición de las ciencias físicas que permitiría demostrar la estructura cuantitativa de los atributos psicológicos, la teoría de la *Medición Conjunta (Conjoint*

Measurement). En general, la teoría de la Medición Conjunta se aplica cuando el ordenamiento de un atributo dependiente P cambia con el efecto conjunto de dos atributos independientes X e Y (Heene, 2013; Perline, Wright y Wainer, 1979).

Así como la estructura cuantitativa que defiende Michell (1999), la Medición Conjunta también postula axiomas que delimitan la estructura cuantitativa de los tres atributos implicados (presentados en la tabla 2.5). Sin embargo, esta aproximación es más compleja que las teorías previamente mencionadas y una descripción completa de los axiomas puede ser consultada en la obra *Simultaneous Conjoint Measurement* (Luce y Tuckey, 1964).

Tabla 2.5

Los requerimientos de la Medición Conjunta

N	Requerimiento
I	La variable P posee un número infinito de valores
II	$P = f(A, X)$ en donde f es una función matemática
III	Existe un orden simple, $\geq$, entre los valores de P
IV	Los valores de A y X pueden identificarse (los objetos pueden ser clasificados de acuerdo a un valor de A y un valor de X)
V	P, A y X son cuantitativos
VI	f es una función no interactiva

Nota. Adaptado de *An introduction to the logic of psychological measurement* (p. 69), por J. Michell, 2014, New York, NY; Psychology Press. Copyright 1990 por Lawrence Erlbaum Associates.

Si se satisfacen los requerimientos I, II, III y IV de la tabla 2.5, entonces la relación entre P, A y X describe un *sistema conjunto*. Para que este sistema cumpla con los requerimientos V y VI, se deben cumplir tres condiciones: *doble cancelación*, que supone que para cada par de valores de P ordenados, entonces otro par de valores también se encontrarán ordenados por la misma

relación (ver figura 2.1); *solubilidad*, que indica que para cada valor de A, existe un valor correspondiente de X; y la *condición arquimédica*, que dicta que toda diferencia entre los valores de un atributo no puede ser infinitamente menor o mayor que otra diferencia; entonces, dadas dos diferencias entre valores del atributo A (o X), la diferencia menor multiplicada por un número natural debe ser por lo menos igual a la diferencia mayor (ver figura 2.2; Michell, 1999).

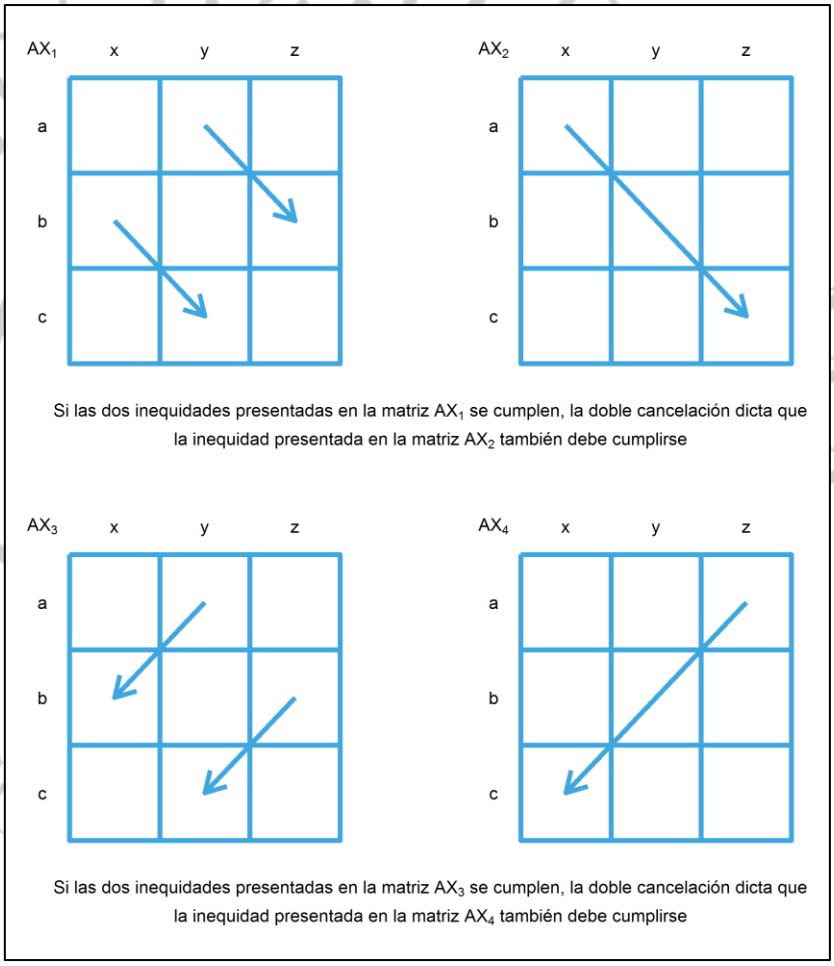

Figura 2.1. La ilustración de la condición de doble cancelación en matrices. Si a, b, c representan valores ordenados de A; x, y, z representan valores ordenados de X. La combinación entre los valores de A y X produce una matriz de relaciones. En la matriz AX_1, se cumple que $(a, y) > (b, z)$ solo si $a + y > b + z$. También, $(b, x) > (c, y)$ solo si $b + x > c + y$. La doble cancelación se satisface si al cancelar los términos comunes de ambas ecuaciones se obtiene que $(a, x) > (c, z)$, representado en la matriz AX_2. En otras palabras, las inequidades presentadas en las

matrices AX_1 y AX_2 no deben contradecirse. El mismo caso ocurre en las matrices AX_3 y AX_4, pero con una inequidad en una dirección opuesta. Adaptado de "The Rasch Model as Additive Conjoint Measurement," por R. Perline, B. D. Wright y H. Wainer, 1946, *Applied Psychological Measurement, 3*(2), p. 242. Copyright 1979 por West Publishing.

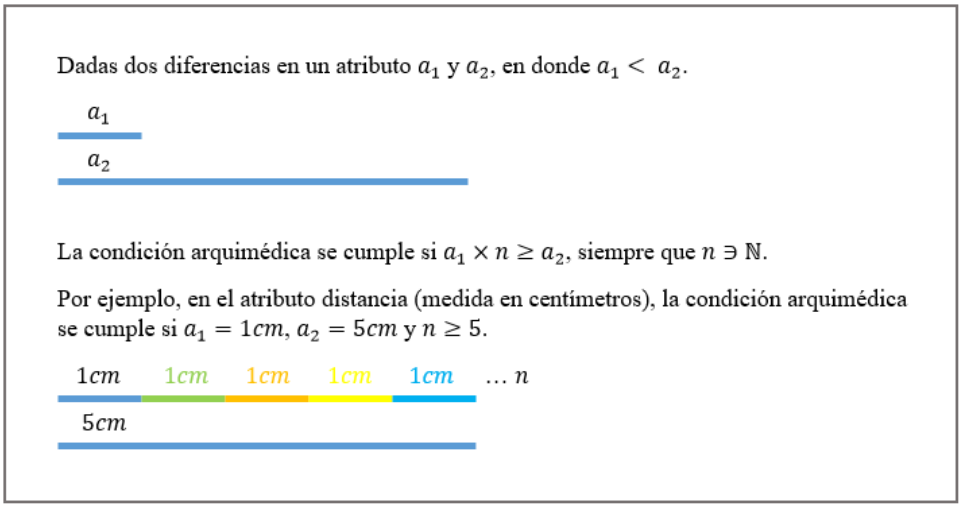

Figura 2.2. Un ejemplo de la condición arquimédica en la Medición Conjunta. Toda diferencia entre los valores de un atributo no puede ser infinitamente menor o mayor que otra diferencia. Entonces, la diferencia menor debe ser por lo menos igual o mayor a la diferencia mayor al ser multiplicada por un número natural. Si se considera al atributo distancia (en cm), 1cm multiplicado por el número natural 5) demuestra el cumplimiento de la condición arquimédica para dicho atributo porque el iguala ambas diferencias expresadas en centímetros.

Michell (2000; 2002; 2019) avala la teoría de la medición conjunta; incluso, la considera como una revolución científica que permitiría producir mediciones en psicología con los mismos estándares que en las ciencias físicas. No obstante, con la excepción de algunos estudios realizados sobre la base de esta teoría (e.g., Kahneman y Tversky, 1979; Karabatsos, 2017; Perline et al., 1979), la práctica psicométrica actual permanece impregnada por la influencia del modelo de Stevens (1946). Por estos motivos, Cliff (1992) denomina a la teoría de la Medición Conjunta como "la revolución que nunca sucedió" (p. 186).

En este contexto, el modelo Rasch surge como una aproximación a la Medición Conjunta de Luce y Tukey (1964). En este modelo de medición se establece que la probabilidad de acertar a un ítem se encuentra determinada por la diferencia entre la habilidad de las personas y la dificultad de los ítems (Wright y Stone, 1979). Esta formulación es estructuralmente equivalente con la Medición Conjunta, pues la probabilidad de acierto representa a un atributo dependiente que es una función aditiva de la habilidad de las personas y la dificultad de los ítems, dos atributos independientes (Borsboom y Scholten, 2008).

Dada la equivalencia estructural entre ambas perspectivas, algunos autores consideran que el modelo Rasch es un caso especial de la teoría de Medición Conjunta (Borsboom, y Scholten, 2008; Perline et al., 1979). Mientras que otros niegan esta posibilidad, argumentando que el modelo Rasch asume que los atributos son de naturaleza cuantitativa; mientras que, la teoría de medición conjunta especifica condiciones ordinales necesarias y/o suficientes para que los tres componentes sean considerados como cuantitativos (Kyngdom, 2008; Michell, 2008; 2014).

Ciertamente, el modelo Rasch no es exactamente una aplicación directa de la teoría de Medición Conjunta, Bond y Fox (2015) reconocen las limitaciones del modelo y; por ello, lo denominan como una *Medición Conjunta Probabilística,* en donde el carácter probabilístico hace referencia a que es imposible determinar con certeza qué sucederá durante la interacción entre una persona y un ítem, pues siempre se encuentra implicado un componente de incertidumbre (Wright y Masters, 1982). Por el contrario, la teoría de la Medición Conjunta es considerada como determinista (Michell, 2007).

A pesar de ello, diversos autores (e.g., Andrich, 1988; Fisher, 1994; Wright, 1985) han demostrado que el modelo Rasch permite obtener mediciones en ciencias sociales cuyas propiedades se asemejan a las medidas de las ciencias físicas. En otras palabras, la evidencia empírica sugiere que este modelo probabilístico permite producir estimaciones con utilidad práctica

con una rigurosidad que se aproxima a los principios que rigen la medición científica (Bond y Fox, 2015).

En síntesis, el modelo Rasch se circunscribe en un panorama histórico de cuestionamientos sobre la posibilidad de realizar mediciones en psicología (Andrich y Marais, 2019). Aunque el modelo no implica una respuesta concreta y consensuada ante las limitaciones de la medición en las ciencias del comportamiento (Borsboom y Scholten, 2008), sus postulados representan una aproximación más cercana a los principios que rigen la medición en las ciencias físicas (Bond y Fox, 2015).

2.2 El modelo Rasch para ítems dicotómicos

George Rasch, matemático danés, propuso en 1960 un modelo de medición para ítems dicotómicos, en donde la probabilidad de acertar a un ítem se encuentra gobernada *únicamente* por la diferencia entre la habilidad de las personas y la dificultad de los ítems (Andrich y Marais, 2019; Bond y Fox, 2015; Kline, 2015). A pesar de la relativa simplicidad del modelo Rasch, diversos autores afirman que las medidas derivadas de dicho modelo poseen propiedades especialmente útiles para las ciencias sociales (Andrich y Malais, 2019; Bond y Fox, 2015; Cavanagh y Waugh, 2011; Waugh, 2007).

Antes de describir detalladamente las características de esta aproximación, es importante reconocer que el modelo Rasch para ítems dicotómicos forma parte de un conjunto de modelos estadísticos denominados *modelos de variables latentes*. Una *variable latente* representa un fenómeno no observable al cual se le atribuye una relación causal frente a un conjunto de indicadores observables (Borsboom, et al., 2003). En este sentido, el objetivo de dichos modelos estadísticos es representar apropiadamente la relación causal de modo que permita describir y predecir el fenómeno no observable de estudio (Bollen y Bauldry, 2011). En el modelo Rasch, los indicadores observables son las respuestas a los ítems, las cuales brindan información de naturaleza ordinal necesaria y suficiente para estimar la

habilidad latente de las personas y dificultad de los ítems (Wright y Douglas, 1986).

Originalmente, Rasch introdujo el modelo para ítems dicotómicos a partir de los siguientes tres supuestos (von Davier, 2016). Primero, que la probabilidad de que la persona p acierte al ítem i es:

$$\Pr\{U_i = 1|p\} = \frac{\lambda_{pi}}{1 + \lambda_{pi}}$$

En donde $U = U_1, \dots, U_I)$ representa un vector de variables observadas discretas y $U_i \in \{0,1\}$; 1 representa un acierto y 0 indica un fallo. De esta manera, para cada evaluado $p = 1, \dots, P$, $U_{pi} \in \{0,1\}$ representa su respuesta en términos de acierto o fallo al ítem i.

Del mismo modo, la probabilidad de que la persona p falle a un ítem $U_i = 0$, puede expresarse como:

$$\Pr\{U_i = 0|p\} = \frac{1}{1 + \lambda_{pi}}$$

Segundo, se cumple que el parámetro $\lambda_{pi} = \frac{\Pr\{U_i=1|p\}}{\Pr\{U_i=0|p\}}$ y puede expresarse como:

$$\lambda_{pi} = \frac{\tau_p}{\xi_i}$$

Lo que implica que $\Pr\{U_i = 1|p\} = \frac{\tau_p}{\xi_i+\tau_p}$ y $\Pr\{U_i = 0|p\} = \frac{\xi_i}{\xi_i+\tau_p}$

Tercero, se asume la independencia estocástica de las probabilidades de la persona p al enfrentarse a múltiples ítems $i = 1, \dots, I$.

Los parámetros τ_p y ξ_i son cantidades desconocidas que describen el nivel de habilidad de la persona p y el grado de dificultad del ítem i, respectivamente (von Davier, 2016). En la notación actual, los parámetros del modelo son transformados como $\theta_p = \ln(\tau_p)$ y $\beta_i = \ln(\xi_i)$ para modificar la expresión a una notación más próxima a los modelos de la Teoría de Respuesta al Ítem [TRI].

$$Pr(U_i = u|p) = \frac{\exp(u(\theta_n - \beta_i))}{1 + \exp(\theta_n - \beta_i)}$$

Que es equivalente a la notación presentada por Bond y Fox (2015):

$$P_{ni}(x_{ni} = 1|\theta_n, \beta_i) = \frac{e^{(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}$$

En donde $P_{ni}(x_{ni} = 1|\theta_n, \beta_i)$ es la probabilidad de acierto $(x = 1)$ de la persona n en el ítem i condicionado por la habilidad θ de la persona n y la dificultad β del ítem i. Esta probabilidad es igual a la constante e o función logarítmica natural (2.7183) elevada a la diferencia entre la habilidad de la persona y la dificultad del ítem $(\theta_n - \beta_i)$ y luego dividida entre 1 más este mismo valor. Para una revisión más extensa sobre el planteamiento del modelo, revisar Andrich (2004) o von Davier (2016).

En cierta forma, el modelo Rasch puede considerarse como un modelo de efectos mixtos y, específicamente, como una regresión logística con efectos mixtos (Mair, 2018). Una regresión logística es básicamente una función que permite determinar la relación entre un conjunto de variables independientes y una variable dependiente de naturaleza categórica con dos posibles resultados. Este modelo se desarrolla sobre la base de la función logística, expresada como

$$f(x) = \frac{1}{(1 + e^{-x})}$$

Entre las propiedades de esta función, los valores de x pueden variar entre $-\infty$ a $+\infty$, es decir, es asintótica; además, el rango de $f(x)$ va de 0 a 1; en consecuencia, es posible modelar la probabilidad de que un evento aleatorio con dos posibles resultados ocurra porque todos los valores se encontraran dentro de este rango (Kleinbaum y Klein, 2010). Finalmente, otra de las propiedades de la función es que, cuando los valores de x empiezan en $-\infty$ y comienzan a ascender, los valores en $f(x)$ incrementan, pero siguen cercanos a cero. No obstante, llega un punto entre los valores de x en el que el incremento en $f(x)$ mejora drásticamente y, conforme se acerca a uno, este incremento continúa, pero en menor grado (ver figura 2.3; Harrell,

2015). Esta característica particular refleja la relación monotónica entre x y $f(x)$, y puede comprobarse al graficar dicha función y observar la forma S o también denominada *sigmoidal* de la curva logística (Peng, Lee y Ingersoll, 2002).

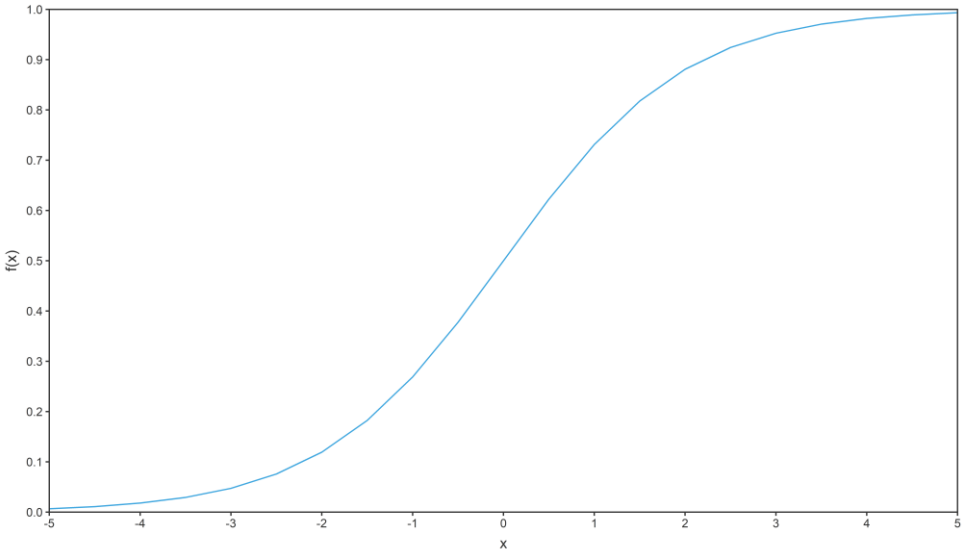

Figura 2.3. La función logística. Aunque los valores de x se encuentren entre -5 y +5, la escala total oscila entre $-\infty$ a $+\infty$.

A diferencia de la regresión lineal, la aproximación logística utiliza el *logit* o *logaritmo natural de los odds ratio* de la variable dependiente.

$$Logit(Y) = natural\log(odds(Y)) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

En donde Y es la variable independiente, π es la probabilidad de que ocurra un determinado evento, β es coeficiente de regresión y α es el intercepto de Y. En el ejemplo de una sola variable predictora, si se utiliza el *antilog* de esta ecuación es posible derivar una ecuación que predice la ocurrencia del resultado de interés 1 como:

$$P(Y = 1|X = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

El valor del coeficiente β determina la dirección de la relación entre X y el logit de Y. Por lo tanto, cuando $\beta > 0$ la relación entre las variables es directa; mientras que, cuando $\beta < 0$, la relación es inversa. Además, β representa el cambio en el logit de Y por unidad que aumenta en X (Peng et al., 2002).

De acuerdo a Mair (2018), el modelo Rasch puede considerarse como un modelo de regresión logística en donde los ítems representan efectos fijos, es decir, constantes y las personas son efectos aleatorios. Los interceptos aleatorios corresponden a los parámetros de personas y los parámetros de efectos fijos corresponden a los parámetros de localización de los ítems. Sin embargo, es importante mencionar que es posible emplear otras funciones de enlace para los modelos de Teoría de Respuesta al Ítem (e. g., probit).

2.2.1 Las ventajas del modelo Rasch frente a la Teoría Clásica

El modelo Rasch para ítems dicotómicos y otros modelos logísticos de la Teoría de Respuesta al Ítem son postulados en un contexto de severas críticas hacia el modelo de medición de la Teoría Clásica de los Tests [TCT]. Entre todas las limitaciones del modelo clásico pueden identificarse cuatro ejes principales a los cuales el modelo Rasch presenta una alternativa idónea de solución.

En primer lugar, se asume sin justificación alguna que la suma de las puntuaciones independientes de un ítem o *parcel* puede considerarse como una variable en un nivel de medición de intervalo (Stevens, 1946; Streiner, 2010; ter Laak, Gokhale y Desai, 2013). La propia suma de respuestas a ítems es cuestionable debido a que la naturaleza categórica de dichas respuestas imposibilita toda combinación lineal por el simple hecho de que los numerales asignados no representan cantidades, sino numerales cuya única propiedad es denotar orden (Stevens, 1946; Bond y Fox, 2015).

En contraste, el resultado de un análisis a través del modelo Rasch es una transformación logarítmica de datos ordinales que deriva en una escala de intervalo común para la habilidad de las personas y la dificultad de los

ítems, una de las propiedades más útiles del modelo (Andrich y Marais, 2019). El nivel de medición de intervalo de los parámetros transformados supone la posibilidad de hacer inferencias en relación a la distancia entre los numerales asignados (Bond y Fox, 2015). De esta manera, esta propiedad posibilita el uso de análisis estadísticos más refinados cuyos supuestos implican variables cuantitativas (Stevens, 1946).

Esta métrica común entre la habilidad de las personas y la dificultad de los ítems tiene como unidad de medida el *logit* (*log odds unit*). Para fijar los valores de esta nueva escala, se suele establecer la media del parámetro de dificultad de los ítems con el valor de 0.0 logits (Wright y Stone, 1979). A partir de esta escala es posible ordenar a personas según su nivel de habilidad y a ítems según su grado de dificultad (Wright y Masters, 1982). Esta propiedad brinda información importante para el proceso de diseño y construcción de instrumentos, especialmente en el contexto de las evaluaciones adaptativas computarizadas.

En segundo lugar, las medidas obtenidas a través de la TCT presentan dependencia frente al test y a la muestra de personas (Paek y Cole, 2020). Por el contrario, la formulación matemática del modelo Rasch posee propiedades que permiten demostrar la *invarianza*, que supone que la dificultad de los ítems puede ser estimada independientemente de las personas que son evaluadas y; del mismo modo, la habilidad de las personas puede ser estimada independientemente de los ítems a los que se enfrentan (Bond y fox, 2015, Waugh, 2007).

No obstante, es importante aclarar que la propiedad supone una *invarianza de la comparación* u *ordenamiento* entre ítems y personas, no necesariamente de la estimación de sus parámetros (Andrich y Marais, 2019). Efectivamente, al revisar nuevamente la ecuación presentada por Rasch, si los datos obtenidos de la interacción entre todas las personas p de una población y todos los ítems $i = 1, \dots, I$ se ajustan al modelo Rasch, es posible demostrar que la comparación de personas puede realizarse de manera independiente a los ítems involucrados.

Para cada ítem i, el ratio de los parámetros de dos personas p y p' equivale a $\frac{\lambda_{ip}}{\lambda_{ip'}} = \frac{\tau_p}{\tau_{p'}}$ independientemente de qué ítem sea escogido para dicha comparación. De manera similar, las comparaciones entre los ítems i e i' puede realizarse independientemente de las personas de referencia, pues para cualquier evaluado p, la comparación entre ítems equivale a $\frac{\lambda_{ip}}{\lambda_{i'p}} = \frac{\xi_i}{\xi_{i'}}$ (von Davier, 2016).

Finalmente, otra de las ventajas de los modelos de Teoría de Respuesta al Ítem y del modelo Rasch frente a la TCT es la relativa facilidad para realizar una equiparación de medidas (von Davier, 2016). En efecto, una comparación directa entre dificultades de ítems o niveles de habilidad de personas es posibles cuando dos conjuntos de observaciones se superponen parcialmente, es decir, cuando ambos conjuntos tienen ítems o personas en común. Por ejemplo, si una persona p fue evaluada solo con los ítems i e i'; mientras que la persona q fue evaluada a partir de los ítems i' e i'', la comparación de los ítems i e i'' puede realizarse si los parámetros λ_{ip} son conocidos (von Davier, 2016). Esto ocurre porque se cumple que

$$\frac{(\lambda_{ip}/\lambda_{i'p})}{(\lambda_{i''p}/\lambda_{i'p})} = \frac{(\xi_{i'}/\xi_i)}{(\xi_{i'}/\xi_{i''})} = \frac{(\xi_{i''})}{(\xi_i)}$$

En esta expresión se demuestra que dos ítems que nunca fueron evaluados en el mismo test pueden ser comparados si existe un ítem en común i' que interactuó con un grupo de personas, de modo que es posible vincular a los ítems i e i''. Este mismo procedimiento puede realizarse para comparar personas que fueron evaluadas a través de formas distintas de un test (von Davier, 2016).

2.2.2 La estimación de parámetros en el modelo Rasch

Como se mencionó anteriormente, los parámetros de ítems y personas son desconocidos y deben ser estimados a partir de las respuestas a los ítems (Bond y Fox, 2015). Una aproximación relativamente simple para realizar esta estimación consiste en calcular las tasas de acierto para personas e ítems a través del cociente entre el número de aciertos sobre el número de intentos. Luego, estas tasas de acierto son convertidas a *odds*, una razón de éxito y fallo. Finalmente, esta razón puede ser transformada a través de una función logarítmica natural (Bond y Fox, 2015).

Por ejemplo, en un instrumento compuesto por 10 ítems, si una persona acertó a 9 de ellos; entonces su tasa de acierto es de 90%, lo que corresponde a una razón de acierto y fallo de 90/10, cuya transformación logarítmica natural es 2.197. La estimación de los parámetros de dificultad de los ítems requiere de un procedimiento similar, se utiliza una transformación logarítmica de la razón acierto y fallo de la tasa de aciertos de un ítem (Bond y Fox, 2015). Como resultado, las estimaciones de los parámetros son expresadas en una escala de *log odd ratios* o *logits*.

Este ejemplo es utilizado particular es utilizado por Bond y Fox (2015) para describir de manera sencilla la idea general del proceso de estimación de parámetros. No obstante, en la literatura metodológica se presentan distintas aproximaciones para realizar dicha estimación; entre ellas, algunas permiten estimar ambos parámetros de manera simultánea; mientras que, otros se limitan a la estimación única de parámetros de ítems o personas (Hambleton y Swaminathan, 1991). En la tabla 2.6 se presentan los principales métodos de estimación en el contexto de modelos Rasch.

Tabla 2.6

Clasificación de los métodos de estimación de parámetros

Categoría	Método	Descripción
Estimación simultánea de parámetros de ítems y personas	*Joint maximum likelihood estimation* [JMLE] (Birnbaum, 1968)	JMLE es el método más popular para estimar parámetros del modelo Rasch; consiste en un proceso iterativo entre la calibración de ítems y medición de personas que considera a ambos parámetros como efectos fijos.
	Bayesian methods [BAYES]	Los métodos BAYES pertenecen a la perspectiva estadística bayesiana, en donde la información previa sobre la distribución de los parámetros a estimar es utilizada en conjunto con los datos empíricos para estimar con mayor precisión a los parámetros.
Estimación de parámetros de ítems	*Conditional Maximum Likelihood Estimation* [CMLE] (Andersen, 1973)	CMLE fue uno de los métodos sugeridos por Rasch (1960), que consiste en utilizar los puntajes brutos de las personas como una representación suficiente de los parámetros de habilidad, de modo que no sean requeridos para la calibración de los ítems.
	Marginal Maximum Likelihood Estimation [MMLE] (Bock y Lieberman, 1970)	MMLE es uno de los métodos más utilizados en la estimación de parámetros, su característica principal es asumir un supuesto distribucional (usualmente una distribución normal) sobre la distribución de la variable latente, tratándola como un efecto aleatorio. De esta manera, remueve a las personas del proceso de calibración de ítems.
Estimación de parámetros de personas	*Maximum Likelihood Estimation* [MLE] (Birbaum, 1968; Lord, 1980).	Utiliza un proceso iterativo para estimar parámetros de personas. La metodología es similar al JMLE, pero los parámetros de ítems son establecidos a priori. Asigna parámetros a personas de modo que se maximice la verosimilitud de los datos observados

	enfocando las estimaciones en la moda de la función de verosimilitud.
Weighted Likelihood Estimation [WLE] (Warm, 1989)	El método surge como respuesta ante el sesgo en las estimaciones del MLE; proponiendo una función de verosimilitud ponderada para corregir el sesgo enfocando las estimaciones en la media de la función.
Maximum A-Posteriori [MAP] (Samejima, 1969)	Método bayesiano con una metodología equivalente al MLE, pero incluyendo una distribución previa de los parámetros de personas. Si la distribución es uniforme, debería generar los mismos parámetros que el MLE. El valor del parámetro de habilidad asignado a cada persona es la moda de su distribución posterior.
Expected A-Posteriori [EAP] (Bock y Aitkin, 1981; Bock y Mislevy, 1982)	A diferencia del MAP, este estimador utilizar la media de la distribución posterior de cada persona para asignar un parámetro de habilidad.

Nota. Además de los métodos presentados en esta tabla existen una amplia variedad de métodos de estimación en el contexto de los modelos Rasch. Para una revisión exhaustiva, revisar Linacre (1999) y Bock y Mislevy (1982). Adaptado de *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments* (pp. 238-251), por G. Engelhard y S. A. Wind, 2018, New York, NY: Routledge. Copyright 2018 por Taylor & Francis.

2.2.3 Supuestos del modelo Rasch

El modelo Rasch para ítems dicotómicos establece un marco de medición sobre la base de un principio fundamental, el cual Rasch (1960/1980) define de la siguiente manera:

> Una persona con más habilidad que otra tendrá una mayor probabilidad de acierto en todos los ítems y; similarmente, un ítem con más dificultad que otro significa que, para cualquier persona, la probabilidad de acertar al segundo será mayor (p. 117).

Este principio, aunque parezca bastante simple, involucra una serie de supuestos que delimitan la teoría de medición que subyace al modelo Rasch (Wright y Stone, 1979). Además, de este mismo enunciado se presentan tres condiciones que resultan del enfrentamiento entre una persona con un determinado nivel de habilidad y un ítem (Andrich y Marais, 2019):

$$si\ \theta_n > \beta_i, entonces\ P(x_{ni} = 1|\theta_n, \beta_i) > 0.5$$

$$si\ \theta_n < \beta_i, entonces\ P(x = 1|\theta_n, \beta_i) < 0.5$$

$$si\ \theta_n = \beta_i, entonces\ P(x = 1|\theta_n, \beta_i) = 0.5$$

En la primera ecuación, si el parámetro de habilidad de la persona θ_n es *mayor* que el parámetro de dificultad del ítem β_i, entonces la probabilidad de acierto será mayor a 0.5. Mientras que, en la segunda ecuación se establece que si el parámetro de habilidad de la persona θ_n es *menor* que el parámetro de dificultad del ítem β_i, entonces la probabilidad de acierto será menor a 0.5. Finalmente, si el parámetro de habilidad de la persona θ_n es igual al parámetro de dificultad del ítem β_i, entonces la probabilidad de acierto será igual a 0.5 (Andrich y Marais, 2019).

El cumplimiento de estas condiciones resulta en las propiedades de invarianza de la comparación entre medidas del modelo Rasch y la estructura aditiva que permite realizar operaciones de concatenación y análisis estadísticos complejos (Bond y Fox, 2015). No obstante, para obtener dichas propiedades, los datos deben adherirse a una serie de supuestos que subyacen al modelo Rasch (Andrich y Marais, 2019). En la tabla 2.7 se presentan los principales supuestos del modelo.

Tabla 2.7

Supuestos del modelo Rasch

Supuesto	Descripción
Unidimensionalidad	Una variable latente dominante causa las covariaciones entre las variables observadas. En otras palabras, las respuestas a los ítems se encuentran determinadas por un solo atributo subyacente (Kingston, Scheuring y Kramer, 2013). Linacre (2011) afirma que las medidas Rasch siempre cumplen con la unidimensionalidad, pues esta es una propiedad forzada por el propio modelo y verificada a través de los índices de ajuste.
Monotonicidad	La probabilidad de acertar a un ítem y la habilidad de la persona tienen una relación monotónica, es decir, incrementan o disminuyen simultáneamente (Reise, Moore y Haviland, 2013).
Independencia local	Los ítems empleados para un análisis Rasch deben ser independientes. En otras palabras, el desempeño de una persona en un ítem no debe afectar su respuesta a cualquier otro ítem del test (Hambleton y Swaminathan, 1991). La única relación que existe entre ellos es a través de la variable latente subyacente. En consecuencia, si se controla el efecto de dicha variable, no debería existir una relación entre las respuestas a los ítems o los residuales (Baghaei, 2008).
Homogeneidad	Para cualquier valor de la habilidad de la persona, el ordenamiento de los ítems en términos de probabilidades será el mismo. Por ejemplo, el ítem más fácil de un test será el ítem más fácil para todos los participantes (Christensen et al., 2013).
Ausencia de funcionamiento diferencial de los ítems [FDI]	El FDI es reconocido como una fuente potencial de sesgo en la medición de personas (Tennant y Pallant, 2007). En general, se refiere a la situación en donde dos individuos de distintos subgrupos de la población con el mismo nivel en el rasgo latente tienen probabilidades distintas de acertar a un ítem (Rogers y Swaminathan, 2017).

Nota. Muchos de estos supuestos también son compartidos por otros modelos de Teoría de Respuesta al Ítem. Adaptado de *Rasch models in health* (p. 11), por K. G. Christensen, S. Kreiner y M. Mesbah, 2013, Hoboken, NJ: John Wiley & Sons. Copyright 2013 por ISTE.

2.2.4 Ajuste del modelo Rasch

Para determinar si los ítems o las personas son congruentes con este marco unidimensional y las propiedades que delimita el modelo, existen dos estadísticos de ajuste denominados *infit* y *outfit* (Bond y Fox, 2015; Linacre y Wright, 1994). Su cálculo se realiza sobre la base de los residuos cuadráticos estandarizados del modelo Z_{ni}^2, los cuales se estiman a través de la siguiente fórmula:

$$Z_{ni}^2 = \frac{R_{ni}^2}{VAR(x_{ni}^2)}$$

En donde R_{ni}^2 representa los residuales elevados al cuadrado, estos se definen como la diferencia entre el valor observado con el valor esperado por el modelo Rasch; y $VAR(x_{ni}^2)$ es la variabilidad de las respuestas observadas.

El *infit* (information weighted fit statistic) consiste en asignar mayor ponderación al desempeño de las personas ubicadas cerca al valor de dificultad del ítem y; por lo tanto, es más sensible a respuestas inesperadas en ítems calibrados cerca de la medida de las personas. Puede calcularse a partir de la siguiente fórmula:

$$INFIT_i = \sum_{N}^{n=!} w_{ni} Z_{ni}^2$$

El *outfit* (outlier sensitive fit statistic) no realiza ninguna ponderación; por ello es más sensible a valores extremos que, en este contexto son respuestas inesperadas a los ítems más fáciles o más difíciles de toda la escala.

$$OUTFIT_i = \frac{1}{N} \sum_{N}^{n=!} Z_{ni}^2$$

Los valores de estos índices indican la presencia de ruido en los datos. Al ser derivados de las diferencias cuadráticas, su valor esperado es 1; por ello, los valores cercanos a 1 implican menor presencia de ruido en los datos. Del mismo modo, los valores por debajo de 1 indican observaciones muy predecibles, redundancia o sobreajuste al modelo; mientras que, los valores mayores a 1 indican impredictibilidad, ruido no modelado o desajuste al modelo (Linacre 2002). En la tabla 2.8 se presentan las recomendaciones de Linacre (2002) para interpretar los valores de los índices de ajuste al modelo Rasch.

Tabla 2.8

Implicancia del valor de los índices de ajuste para la medición

Valor del índice	Implicancia para la medición
> 2.0	Distorsiona o afecta el sistema de medición.
1.5 – 2.0	No productivo para construir mediciones, pero no afecta el sistema.
0.5 – 1.5	Productivo para la medición.
> 0.5	Menos productivo para la medición, pero no afecta al sistema. Puede presentar un sesgo al producir índices de confiabilidad altos.

Nota. Adaptado de "What do Infit and Outfit, Mean-square and Standardized mean?." Por J. M. Linacre, 2002, *Rasch Measurement Transactions, 16*(2), p. 878. Copyright 2002 por Rasch Measurement SIG, AERA.

Adicionalmente, Wright y Linacre (1994) recomiendan que, para delimitar los valores adecuados para ambos estadísticos, es necesario tomar en cuenta el tipo de evaluación que se pretende implementar; por ello, proponen un rango de valores idóneos para casos particulares, detallados en la tabla 2.9.

Tabla 2.9

Rangos aceptables para los índices Infit y Outfit según tipo de evaluación

Rango	Tipo de evaluación
0.8 – 1.2	Preguntas de opción múltiple en pruebas de altas consecuencias.
0.7 – 1.3	Preguntas de opción múltiple en pruebas regulares.
0.6 – 1.4	Escalas de calificación o cuestionarios en general.
0.5 – 1.7	Observaciones clínicas
0.4 – 1.2	Juicios en donde se fomenta un acuerdo

Nota. Adaptado de "Reasonable mean-square fit values" Por B. D. Wright y J. M. Linacre, 1994, *Rasch Measurement Transactions, 8*(3), p. 370. Copyright 1994 por Rasch Measurement SIG, AERA.

Una alternativa adicional a los índices presentados es la correlación entre la alternativa de respuesta correcta de un ítem y la medida estimada para cada persona (MINEDU, 2018). Si un ítem es congruente con el marco propuesto en el modelo Rasch, entonces se espera que dicha relación sea positiva, es decir, a mayor habilidad, mayor será el acierto en el ítem (Linacre, 2019a). Esto se determina a partir de la siguiente fórmula:

$$r_{X\theta} = \frac{\sum_{n=1}^{N}\left(X_n - \sum_{m=1}^{N}\frac{X_m}{N}\right)\left(\theta_n - \sum_{m=1}^{N}\frac{\theta_m}{N}\right)}{\sqrt{\sum_{n=1}^{N}\left(X_n - \sum_{m=1}^{N}\frac{X_m}{N}\right)^2 \sum_{n=1}^{N}\left(\theta_n - \sum_{m=1}^{N}\frac{\theta_m}{N}\right)^2}}$$

En donde X representa la respuesta $(0; 1)$ al ítem n y θ es el parámetro de habilidad de la persona m. Esta estrategia se denomina correlación *Punto Biserial* y representa un caso particular de la correlación de *Producto Momento de Pearson* cuando una de las variables es dicotómica y la otra es continua (Field, 2018). En la práctica, esta correlación se realiza a todas las alternativas de respuesta de un ítem, si la alternativa correcta tienen una correlación positiva significa que las personas con mayor habilidad están acertando al ítem; mientras que las personas con menor habilidad están

fallando. En la mayoría de casos, si esta relación es negativa el ítem desajustará con el modelo, pues esta situación implica que las personas con mayor habilidad están fallando al ítem y las personas con menor habilidad lo están acertando.

2.2.5 Análisis gráfico del modelo Rasch

Las propiedades del modelo Rasch pueden apreciarse con detenimiento al observar las Curvas Características de los Ítems [CCI]. Las CCI son gráficos que contrastan la probabilidad de acierto con la escala logits que puede representar a ítems y personas (Bond y Fox, 2015). En este plano, cada ítem es representado por una curva de probabilidad acumulada cuya localización es determinada por el nivel de dificultad del ítem y el punto de inflexión que indica la probabilidad de 0.5 de acierto (Heise, 2010).

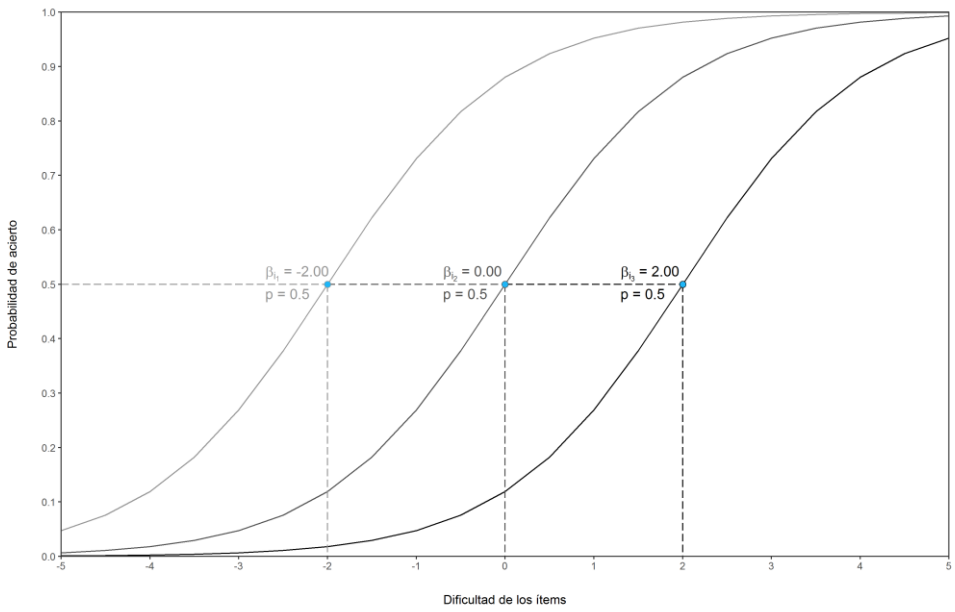

Figura 2.4. Curvas características de tres ítems hipotéticos. El ítem 1 (i_1) es representado por la curva de la izquierda, el ítem 2 (i_2) por la curva central y el ítem 3 (i_3) por la curva de la derecha. A pesar de que el límite de la escala logit en el eje *x* es de -5 y 5, este rango de valores en realidad es asintótico, pues va desde $-\infty$ a ∞. En contraste, la probabilidad de acierto en el eje *y* siempre va desde 0 a 1.

En la figura 2.4, el ítem 1 (i_1) tiene una dificultad de -2.00, el ítem 2 (i_2) tiene una dificultad de 0.00; y el ítem 3 (i_3) tiene una dificultad de 2.00. Este gráfico es particularmente útil para determinar qué ítems tienen mayor dificultad que otros, pues una mayor dificultad implicará siempre un posicionamiento a la derecha. Al mismo tiempo, como todas las curvas poseen la misma pendiente, un ítem más difícil que otro siempre tendrá una menor probabilidad de acierto (Bond y Fox, 2015).

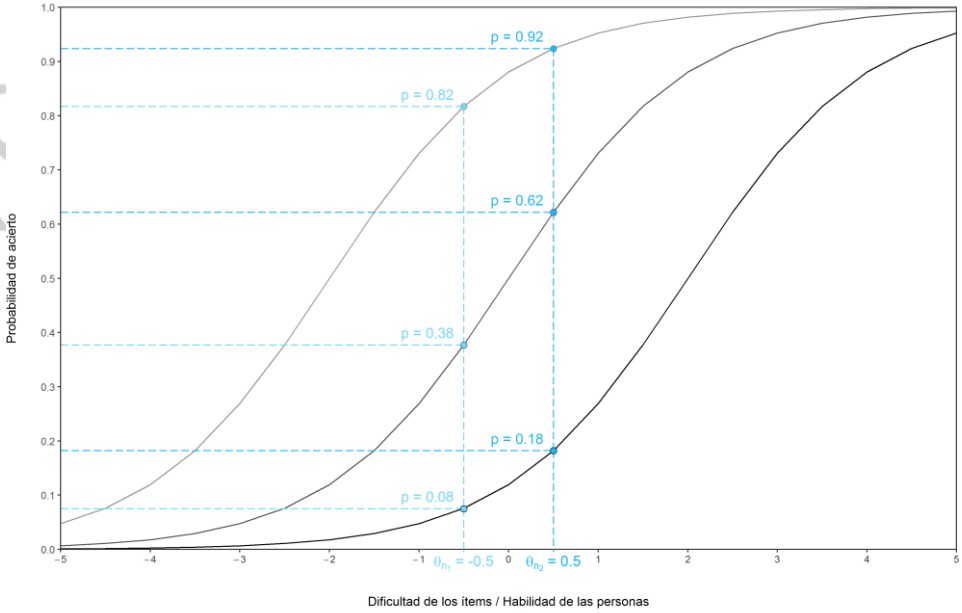

Figura 2.5. Curvas características de tres ítems hipotéticos y su relación con dos personas hipotéticas. La persona 1 (n_1) se representa por la línea punteada de la izquierda. La persona 2 n_2 es representada por la línea punteada de la derecha. A pesar de que el límite de la escala logit en el eje *x* es de -5 y 5, este rango de valores en realidad es asintótico, pues va desde $-\infty$ a ∞. En contraste, la probabilidad de acierto en el eje *y* siempre va desde 0 a 1.

Al incorporar a personas al gráfico en la figura 2.5, es posible determinar sus relaciones con los ítems previamente establecidos; por ejemplo, la persona n_1 tiene una habilidad de -0.5 y; por lo tanto, una probabilidad de 0.82 de acertar al primer ítem, 0.36 para el segundo y 0.08 para el tercero. Del mismo modo, la persona n_2 tiene una habilidad de 0.5; lo que implica

una probabilidad de 0.92 de acertar el primer ítem, 0.62 para el segundo ítem y 0.18 para el tercero. En este sentido, las CCIs permiten evidenciar la premisa principal del modelo, pues toda persona con mayor habilidad que otra siempre tendrá mayor probabilidad de acierto en todos los ítems. Del mismo modo, un ítem con mayor dificultad siempre tendrá menor probabilidad de acierto que ítems con menor dificultad (Bond y Fox, 2015).

Otra alternativa especialmente útil en el diseño y construcción de instrumentos son los *mapas de Wright*, una herramienta gráfica que permite contrastar la distribución de los parámetros de ítems y personas (Callingham y Bond, 2006). La construcción de esta técnica gráfica consiste en establecer la media aritmética de los parámetros de dificultad de los ítems como punto el punto central del gráfico en un punto arbitrario (usualmente en 0.0 logits) y; a partir de ello, cada ítem es ubicado en la zona derecha del gráfico en función a la distancia entre su dificultad y el promedio. En la zona izquierda, cada persona es ubicada en el punto en donde tengan un 50% de probabilidades de responder a un ítem correctamente (Bond y Fox, 2015). En la figura 2.6 se presenta un ejemplo de este gráfico.

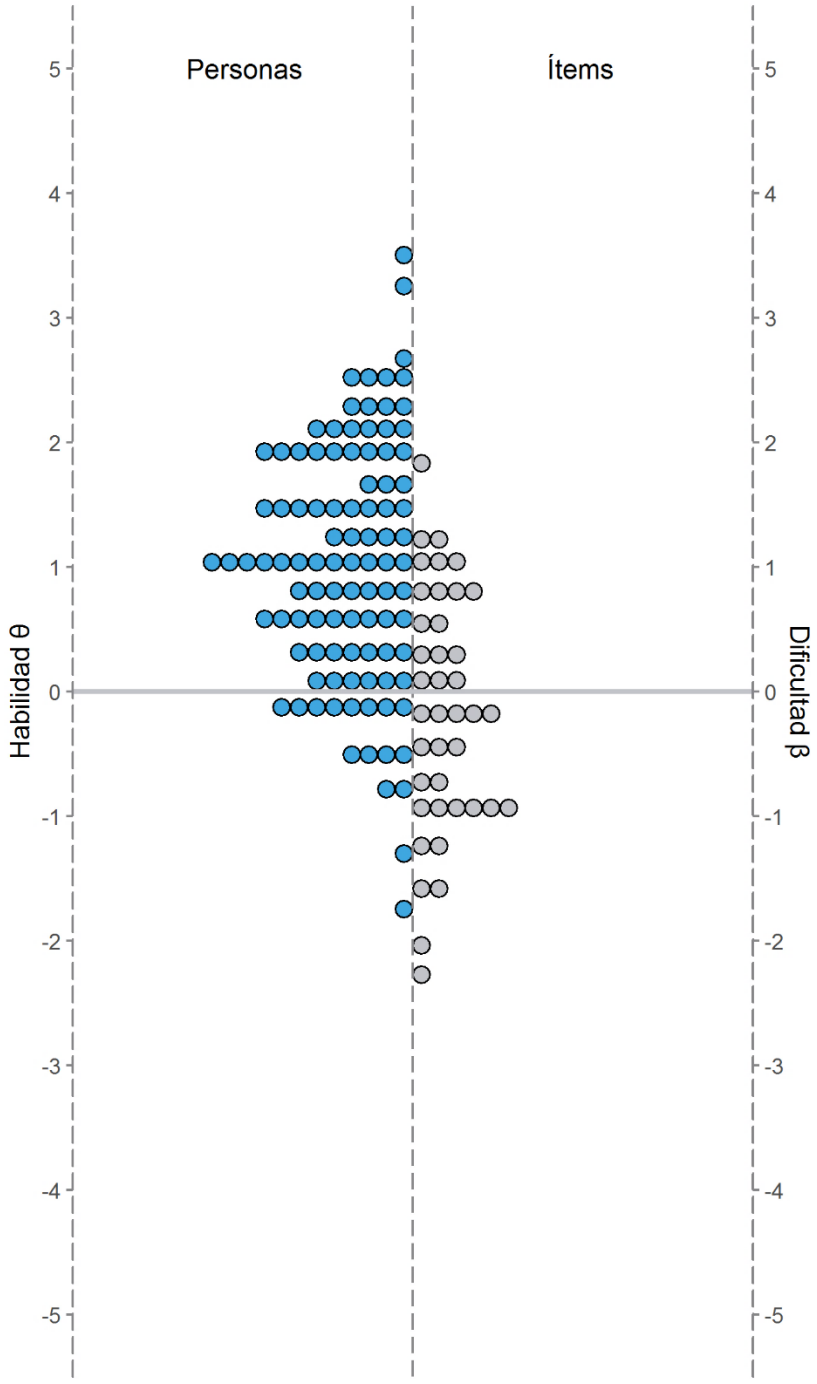

Figura 2.6. El mapa de Wright. Este gráfico fue desarrollado considerando la simulación de 40 ítems y 100 personas con parámetros extraídos a partir de una distribución normal estándar para ambos casos).

2.2.6 El modelo Rasch y otros modelos unidimensionales para ítems dicotómicos

En varias publicaciones académicas, distintos autores consideran al modelo Rasch como el modelo logístico de un parámetro [1PL] dentro del contexto de los modelos unidimensionales para ítems dicotómicos de la Teoría de Respuesta al Ítem (Hambleton y Swaminathan, 1991). En la notación TRI, la dificultad de los ítems se representa por el parámetro b; de esta manera, la ecuación que representa el 1PL es equivalente a la presentada anteriormente, pero añadiendo una constante D que minimiza la diferencia máxima entre la función de la distribución normal y logística (Camilli, 1994; Linacre, 2005):

$$P_{ni}(x_{ni} = 1|\theta_n, b_i) = \frac{e^{D(\theta_n - b_i)}}{1 + e^{D(\theta_n - b_i)}}$$

No obstante, la familia de modelos Rasch se encuentra estrechamente relacionada con la teoría de la medición conjunta probabilística; por lo tanto, estos modelos presentan un marco prescriptivo de medición al cual los datos deben adherirse. En contraste, los modelos TRI son de carácter descriptivo, es decir, buscan ajustarse a los datos empíricos con el objetivo de explicar la mayor cantidad de varianza (Bond y Fox, 2015; Shaw, 1991).

A pesar de estas diferencias, diversos académicos del campo de los modelos de medición TRI sostienen que el modelo Rasch presenta dos suposiciones implícitas que son objeto de constantes críticas (Kline, 2015). En primer lugar, el modelo Rasch para ítems dicotómicos asume la uniformidad de la discriminación de los ítems. En el contexto de los modelos TRI, la discriminación (o parámetro a) se refiere a la capacidad de las categorías de respuesta de los ítems para diferenciar a individuos con distintos niveles del rasgo latente (Penfield, 2013). Dicha característica de los ítems es reconocida e introducida incluida en el análisis a través del modelo TRI logístico de dos parámetros [2PL].

$$P_{ni}(x_{ni} = 1|\theta_n, b_i, a_i) = \frac{e^{Da_i(\theta_n - b_i)}}{1 + e^{Da_i(\theta_n - b_i)}}$$

La fórmula del modelo 2PL es similar a la del 1PL, solo que introduce el parámetro a como un valor que modifica la pendiente de la ecuación (Paek y Cole, 2020). El problema asociado a la inclusión del parámetro a es que resulta en una contradicción al principio fundamental del modelo Rasch (1960), pues en algunas situaciones, un ítem con mayor dificultad no siempre tendrá menor probabilidad de acierto que un ítem con menor dificultad (Bond y Fox, 2015). Un ejemplo de este caso se presenta en la figuras 2.7.

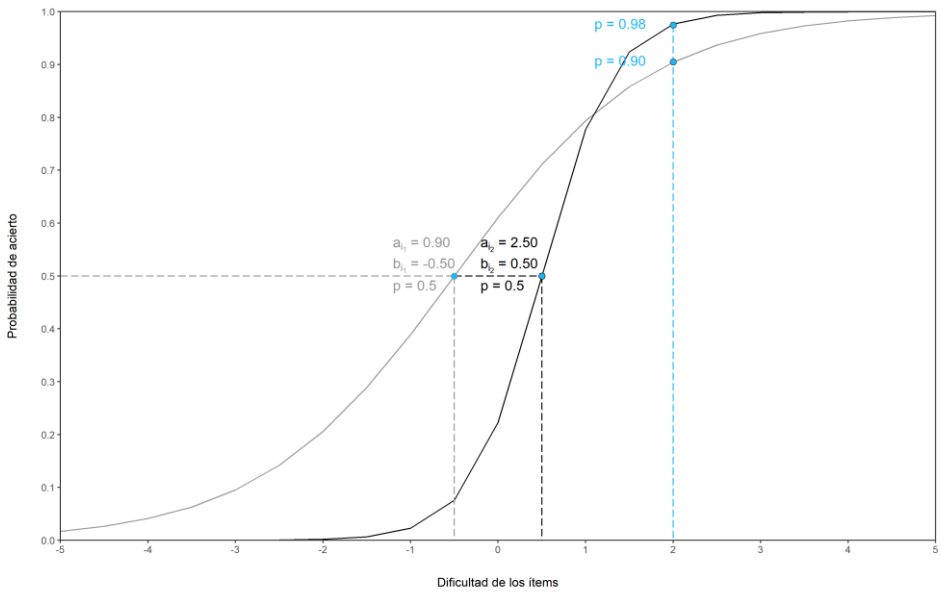

Figura 2.7. Curvas características de dos ítems bajo el modelo 2PL. En este ejemplo, el ítem i_1 tiene una dificultad de $b = -0.5$ y una discriminación de $a = 0.90$; y el ítem i_2 tiene una dificultad de $b = 0.5$ y una discriminación de $a = 2.50$. Dado los postulados del modelo Rasch, el ítem i_1 siempre debería tener una probabilidad de acierto mayor a la del ítem i_2; sin embargo, esto no se cumple en todos los sectores de la curva. Efectivamente, en el valor logit de 2.0, la probabilidad de acierto del ítem i_1 es menor que la del ítem i_2.

En segundo lugar, el modelo Rasch supone la ausencia de la adivinación en las respuestas de los ítems, la cual consiste en la posibilidad de acertar a un ítem por azar. El modelo TRI de tres parámetros [3PL] permite analizar esta característica al añadir un parámetro adicional c que modifica la asíntota inferior de la curva característica del ítem (Paek y Cole, 2020), como se muestra en la figura 2.6.

$$P_{ni}(x_{ni} = 1|\theta_n, b_i, a_i, c_i) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_n - b_i)}}{1 + e^{Da_i(\theta_n - b_i)}}$$

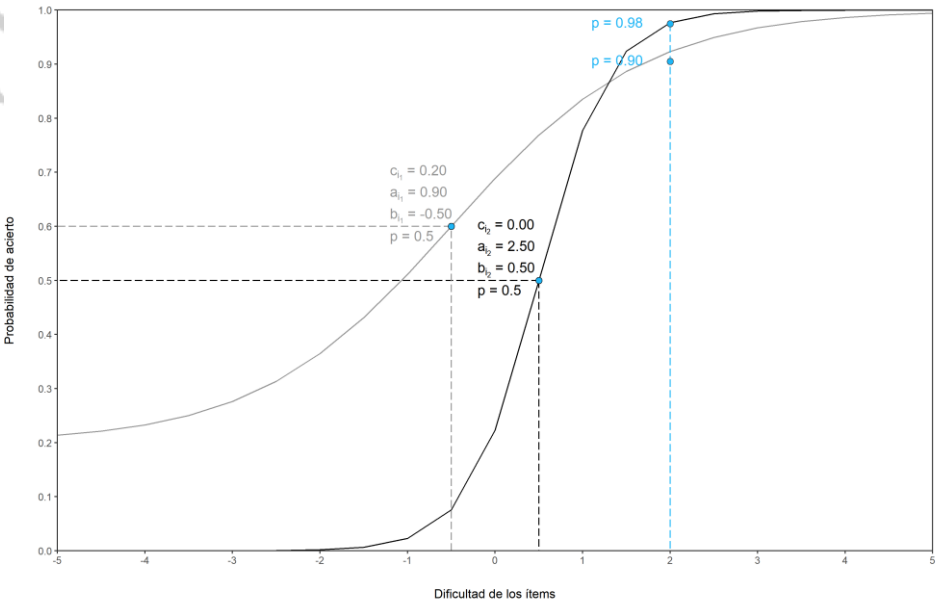

Figura 2.8. Curvas características de dos ítems bajo el modelo 3PL. En este ejemplo, el ítem i_1 tiene una dificultad de $b = -0.5$, una discriminación de $a = 0.90$ y una pseudoadivinación de $c = 0.20$; y el ítem i_2 tiene una dificultad de $b = 0.5$, una discriminación de $a = 2.50$ y una pseudoadivinación de $c = 0.00$. Al igual que en la figura 2.5, no se cumplen los postulados del modelo Rasch, la única diferencia es que el parámetro c modificó la asíntota inferior del ítem i_2, de modo que el parámetro b ya no se ubica en el punto en donde tiene una probabilidad de .50 de acierto, sino en $.50 + \frac{c}{2}$.

Sin embargo, el modelo Rasch no incluye este parámetro en el análisis de ítems ($c = 0$), pues el comportamiento de adivinación es una propiedad de las personas, no de los ítems (Wright y Stone, 1999). Incluso, si existiera una cantidad de adivinación substancial en las respuestas de un ítem particular, este simplemente no se ajustaría al marco de medición que delimita el modelo (Bond y Fox, 2015).

2.3 Confiabilidad

La confiabilidad es considerada como uno de los principios esenciales de la medición psicológica y un requisito indispensable para la validez (Geisinger, 2013). De acuerdo a la AERA et al. (2014), existe una tendencia actual por utilizar el término confiabilidad para referirse únicamente a los coeficientes de confiabilidad de la Teoría Clásica de los Tests [TCT]; sin embargo, la noción moderna de esta propiedad es mucho más extensa.

Tradicionalmente, la noción de confiabilidad se establece sobre la base de un concepto filosófico denominado *puntaje verdadero* (Geisinger, 2013). Este es un componente esencial del modelo de medición establecido en la Teoría Clásica de los Tests, cuya expresión matemática formal se realiza de la siguiente manera:

$$X = T + E$$

El puntaje observado X se compone a partir del puntaje verdadero T y un componente de error E (Desjardins y Bulut, 2018). En esta expresión, el puntaje observado es aquel obtenido en el test; mientras que, el puntaje verdadero es conceptualizado como el valor esperado o el promedio de una distribución hipotética de puntajes que podrían ser obtenidos si un individuo tomara el mismo test un número infinito de veces. Del mismo modo, el error de medición se define como cualquier fluctuación en las puntuaciones observadas que es causada por factores relacionados al proceso de medición y que son irrelevantes para el constructo que se pretende medir (Urbina, 2014).

Desjardins y Bulut (2018) sostiene al utilizar el modelo de la TCT para un proceso de medición, deben considerarse cuatro supuestos adicionales además de la expresión general previamente establecida. En la tabla 2.10 se detallan dichos supuestos.

Tabla 2.10

Supuestos complementarios de la TCT

Supuesto	Descripción
$E(X) = T$	El valor esperado del puntaje observado es el puntaje verdadero.
$Cov(T, E) = 0$	El puntaje verdadero y los errores son independientes.
$Cov(E_1, E_2) = 0$	Los errores entre formas del test son independientes.
$Cov(E_1, T_2) = 0$	Los errores en una forma del test son independientes del puntaje verdadero de otra forma.

Nota. Adaptado de *Handbook of educational measurement and psychometrics using R* (p. 41), por C. D. Desjardins y O. Bulut, 2018, Boca Raton, FL: Taylor & Francis Group. Copyright 2018 por Taylor & Francis Group.

Al incorporar estos supuestos adicionales, el modelo de la TCT puede ser expresado como una simple suma de componentes de varianza σ ortogonales (es decir, no correlacionados):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

En otras palabras, la varianza del puntaje observado σ_X^2 es igual a la suma de la varianza del puntaje verdadero σ_T^2 y la varianza del error σ_E^2. Además, se asume que la varianza del puntaje verdadero es constante, pues no varía

en función a la condiciones como la forma del test o la fecha de aplicación; en contraste, la varianza de error sí fluctúa (Desjardins y Bulut, 2018).

Tomando en consideración lo anterior, la confiabilidad puede definirse como la proporción de la varianza del puntaje observado σ_X^2 que pueda atribuirse a la varianza del puntaje verdadero σ_T^2.

$$confiabilidad = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

De acuerdo a esta expresión, es posible conceptualizar la confiabilidad como el grado en que las puntuaciones se encuentran libres de errores aleatorios (AERA et al., 2014).

Es importante reconocer que toda medición se encuentra asociada a un componente de error, incluso en las ciencias físicas. Por ello, el objetivo de los instrumentos es proveer mediciones lo suficientemente libres de error como para ser útiles (Urbina, 2014). Esta cuestión es particularmente importante para los instrumentos empleados en psicología porque la naturaleza de los constructos evaluados y la metodología empleada implican una mayor susceptibilidad ante distintas fuentes de error (Cohen y Swerdlik, 2010; Kosso, 2011).

En este sentido, la operacionalización de la confiabilidad se encontrará estrechamente relacionada con la fuente de error de medición que pretende cuantificar (Urbina, 2014). En otras palabras, cada coeficiente utilizado para estimar la precisión y consistencia de las puntuaciones dará información relevante sobre una fuente particular de error y un "tipo" de confiabilidad asociado a esta (Desjardins y Bulut, 2018; Geisinger, 2013). En la tabla 2.11 pueden apreciarse los principales tipos de confiabilidad de la TCT en función de la fuente de error que pretenden estimar.

Tabla 2.11

Fuentes de error y tipos de confiabilidad de la TCT

Fuentes de error	Descripción	Tipo de confiabilidad	Método de estimación	Cálculo
Error por muestreo de tiempo	Variabilidad inherente a las puntuaciones de un test en función al hecho de haber sido aplicados en tiempos diferentes. Para ello, se reconoce que el constructo evaluado fluctúa a través del tiempo, incluso aquellos considerados como rasgos relativamente estables como la personalidad.	Coeficiente de estabilidad	Coeficiente de confiabilidad test-retest	Correlación entre las puntuaciones de un test obtenidas en un tiempo determinado con las puntuaciones del mismo test obtenidas en otro intervalo temporal.
Error por muestreo de contenido	Variabilidad que influye en las puntuaciones de un test producto de factores irrelevantes para el constructo medido. Usualmente, surge por deficiencias en el proceso de construcción del test.	Coeficiente de equivalencia	Coeficiente de confiabilidad por formas-alternas	Correlación entre las puntuaciones de un mismo individuo en dos formas distintas de un test.
			Coeficiente de confiabilidad por mitades	Correlación entre las puntuaciones obtenidas al dividir el test en dos partes (esto puede variar en función al tipo de prueba empleado).
Consistencia inter-ítems	Errores que resultan de la fluctuación en los ítems que componen el test.	Coeficiente de consistencia interna	Estadísticos de consistencia interna	Correlación entre mitades equivalentes de un test con la corrección Spearman-Brown. Kuder-Richardson formula 20 [$KR\ 20$] (Kuder y Richardson, 1937) Alfa de Cronbach [α] (Cronbach, 1951)

Nota. Sijtsma (2009) afirma que el método más utilizado para estimar la confiabilidad es el coeficiente α de Cronbach (Cronbach, 1951). Incluso, el trabajo de Cronbach (1951) es uno de los artículos científicos más citados en la literatura (McNeish, 2018; Sijtsma, 2009; van Noorden, Maher y Nuzzo, 2014). No obstante, esta aproximación ha sido objeto de múltiples críticas debido a que su cálculo se basa en una serie de supuestos que difícilmente se cumplen en la realidad (McNeish, 2018; Raykov, 1997; Sijtsma, 2009; Yang y Green, 2011; Zinbarg, Revelle y Yovel, 2005). Por ello, varios autores han propuesto métodos alternativos para la estimación de la confiabilidad como el coeficiente Omega [ω], Omega Total [ω_T], Omega Jerárquico [ω_H] (McDonald, 1999), Omega Total de Revelle [ω_{RT}] (Revelle, 2019), Greatest Lower Bound [GLB] (Jackson y Agunwamba, 1977), entre otros.

La noción moderna de confiabilidad trasciende el marco de la TCT e involucra un espectro más amplio de coeficientes. Dicha noción fue establecida por las instituciones AERA, APA y NCME en los *Standards for Educational And Psychological Testing* o también denominados como *Joint Standards*. De acuerdo a los resultados de varias encuestas realizadas por la International Test Comission [ITC] y la European Federation of Psychological Associations [EFPA], esta publicación es reconocida a nivel internacional como el estándar de calidad técnico y conceptual para la evaluación psicológica y educativa (Bartram y Amado, 2017; Bartram y Hambleton, 2016).

En este documento se define la confiabilidad o precisión como el grado en que los puntajes de un test son consistentes entre réplicas independientes del proceso de medición, independientemente del método empleado para reportar evidencias de dicha consistencia (AERA et al., 2014). Esta idea supone la posibilidad de evaluar la confiabilidad a partir de coeficientes de ajenos a la TCT.

Una de las alternativas para realizar esta evaluación es a través de la Teoría de la Generalizabilidad [Teoría G] (Cronbach, Gleser, Nanda y Rajaratnam, 1972). Dicho modelo es una extensión de la TCT que emplea el análisis de varianza [ANOVA] para evaluar de manera simultánea los efectos combinados de múltiples fuentes de varianza de error en las puntuaciones (Mair, 2018). Sin embargo, el problema con este método es que se requieren múltiples observaciones para cada individuo en todas las variables independientes que puedan contribuir a la varianza de error; por ejemplo, aplicaciones en distintas ocasiones o con formas alternas del test (Urbina, 2014).

Otra alternativa para la estimación de la confiabilidad es a través de la Teoría de Respuesta al Ítem, un conjunto de modelos que tuvieron su origen en los trabajos de Birnbaum, Lord y Rasch entre los años 1950 y 1960. Estas propuestas pretendían superar las limitaciones del modelo de la TCT al modelar la probabilidad de acertar a un ítem a partir de ciertas características de las personas e ítems (van der Linden, 2016). En este

marco, la confiabilidad y el error se estiman a partir de funciones de información de cada ítem, las cuales pueden interpretarse como proposiciones matemáticas acerca de la precisión de la medición en cada nivel del rasgo latente. Sin embargo, estas funciones se basan en los resultados obtenidos en ocasiones específicas y; por lo tanto, no es posible generalizar los resultados a contextos distintos (AERA et al., 2014).

Además de una mayor extensión en relación a los coeficientes para la estimación de la confiabilidad, los estándares también realizaron una declaración explícita sobre este principio. En la práctica, se suele afirmar que la confiabilidad representa una propiedad inmutable del test a través de expresiones como "el test es confiable" o "la confiabilidad del test". El problema con estos enunciados es que suponen que una prueba otorgará resultados consistentes independientemente del contexto en donde se aplique (MINEDU, 2014).

En contraste, la definición moderna designada por la AERA et al. (2014) establece que la confiabilidad es una propiedad de las puntuaciones del test; por lo tanto, se circunscribe en un contexto específico y en un grupo poblacional particular que obtuvo dichas puntuaciones. En consecuencia, es posible obtener distintos grados de confiabilidad de las puntuaciones derivadas de aplicar el mismo instrumento psicométrico en contextos y grupos distintos (Howitt y Cramer, 2017).

Esto se debe a que cuando dos grupos tienen características muy distintas, estas diferencias representan un fuente potencial de error de medición, lo cual se encuentra fuertemente relacionado con el principio de *imparcialidad* de la evaluación psicológica (AERA et al., 2014; Miller, 2013). Dicha cuestión es particularmente problemática para las disciplinas en donde se evalúan poblaciones compuestas por grupos con características muy particulares. En estos casos, es posible que el grado de confiabilidad de toda la población sea adecuado, pero que al analizar cada grupo relevante de manera independiente, existan algunos en donde el grado de confiabilidad sea bajo (Kubiszyn y Borich, 2013; Urbina, 2014).

Por estos motivos, la AERA et al. (2014) recomiendan la estimación de la confiabilidad no solo en toda la población, sino también para cada grupo relevante dentro de ella. Dicha evaluación representa el sustento y enfoque principal del estudio; el cual será abordado a partir de la perspectiva de confiabilidad propuesta sobre la base del modelo Rasch.

2.4 Confiabilidad en el modelo Rasch

Wright y Stone (1999) consideran que la propuesta de Rasch (1960) representa un avance substancial en el campo de la psicometría, pues permite estimar el nivel del rasgo latente en el que se encuentra una persona con mayor precisión que al emplear métodos tradicionales. En congruencia con esta afirmación, diversos autores en este contexto afirman que las medidas obtenidas a partir de este modelo son invariantes e independientes de los ítems utilizados para su estimación (Bond y Fox, 2015; Engelhard y Wind, 2018).

Andrich y Marais (2019) afirman que, en efecto, las medidas de los modelos Rasch suponen cierto grado de invarianza, pero no de la referida a la precisión de la estimación de parámetros, sino a la comparabilidad de ellos. En este sentido, las medidas del modelo Rasch permitirán realizar comparaciones consistentes entre personas y entre ítems. El grado en que dicha invarianza de comparación se cumple representa un indicador de la consistencia de las medidas obtenidas en el proceso evaluación (i.e., confiabilidad).

2.4.1 Coeficientes de confiabilidad del modelo Rasch

La confiabilidad puede ser estimada tanto para los parámetros de habilidad de las personas como para los parámetros de dificultad de los ítems, a través de índices de confiabilidad de separación R (Bond y Fox, 2015; Schumacker y Smith, 2007). En primer lugar, la confiabilidad de separación de personas R_P es una estimación del grado de la replicabilidad del ordenamiento de las personas según su medida de habilidad cuando son

evaluados a partir de otros ítems que miden el mismo constructo (Bond y Fox, 2015).

La estimación de R_P es análoga a la metodología que subyace el alfa de Cronbach (Cronbach, 1951), y se define matemáticamente como la fracción de la varianza observada de las respuestas que puede ser reproducida por el modelo Rasch. Este índice puede expresarse de la siguiente manera:

$$R_p = \frac{SA_p^2}{SD_p^2}$$

En esta expresión, el denominador SD_p^2 representa el total de la variabilidad de personas, es decir, qué tanto difieren las personas en el rasgo de interés. El numerador SA_p^2 representa la parte de esta variabilidad que puede ser reproducida por el modelo Rasch. A esta cantidad reproducible también se le conoce como variabilidad ajustada de personas, la cual se obtiene al substraer la varianza de error SE_p^2 de la varianza total SD_p^2.

$$SD_p^2 - SE_p^2 = SA_p^2$$

La varianza total de los parámetros de habilidad de las personas puede estimarse de la siguiente manera:

$$SD_p^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta_n - \bar{\theta})$$

Asimismo, la varianza de los errores se estima como la media cuadrática de los errores estándar de medición ee.

$$SE_p^2 = \frac{1}{N}\sum_{n=1}^{N} ee(\theta_n)^2$$

En consecuencia, la expresión matemática de la confiabilidad de separación de personas puede definirse de las siguientes formas equivalentes:

$$R_p = \frac{\frac{1}{N-1}\sum_{n=1}^{N}(\theta_n - \bar{\theta}) - \frac{1}{N}\sum_{n=1}^{N} ee(\theta_n)^2}{\frac{1}{N-1}\sum_{n=1}^{N}(\theta_n - \bar{\theta})} = \frac{SD_p^2 - SE_p^2}{SD_p^2} = \frac{SA_p^2}{SD_p^2}$$

La confiabilidad de separación de ítems se estima de la misma manera, pero considerando la variabilidad de los parámetros de dificultad. En ambos casos, los valores de los índices oscilan entre 0 y 1 (Wright y Masters, 1982). Es importante considerar que existen distintos factores que afectan directamente a la estimación de ambos índices (Linacre, 2019a). En la tabla 2.12 se presenta con mayor detalle cada uno de ellos.

Tabla 2.12

Factores que influyen en la estimación de los índices de confiabilidad

Índice de confiabilidad	Factores que incluyen en la estimación	Descripción
Confiabilidad de separación de personas	Variabilidad de la habilidad de la muestra.	Mientras más amplio sea el rango de habilidad, se obtendrán estimaciones mayores.
	Extensión del test (cantidad de ítems).	Mientras más largo es un test, mayor será el valor del índice.
	Número de categorías de respuesta por ítem.	Mientras más categorías de respuesta tengan los ítems, mayor será la confiabilidad.
	Targeting entre muestra e ítems	Un mejor targeting implica una mayor confiabilidad.
Confiabilidad de separación de ítems	Varianza de la dificultad de los ítems	Mientras más amplio sea el rango, mayor será la confiabilidad.
	Tamaño de muestra	A mayor tamaño de muestra, incrementará el valor del índice.

Nota. Adaptado de *A user's guide to WINSTEPS* (p. 671), por J. M. Linacre, 2019 (https://www.winsteps.com/a/Winsteps-Manual.pdf). Copyright 2019 por John M. Linacre.

2.4.1 Targeting

El término targeting hace referencia a una estrategia de ensamblaje de ítems en función del nivel de habilidad de las personas de modo que incremente la precisión de la estimación de las medidas. La metodología de esta estrategia inicia en la delimitación de un *target* de evaluación, el cual puede ser una persona o un grupo de personas. Luego, dado el nivel de habilidad de la persona o la distribución de habilidad del grupo, se generan instrumentos a partir de ítems cuyos niveles de dificultad son congruentes con los niveles de habilidad del target (Wright y Stone, 1979).

La fundamentación que subyace al uso de esta técnica es que permite medir personas con mayor precisión debido a que se emplean ítems *relevantes* para ellas (Jones y Wright, 1992). El carácter de relevancia se refiere al grado en que el ítem aporta información sobre la habilidad de la persona. Un ítem se encuentra perfectamente on target cuando su dificultad es la misma que la habilidad de la persona, de modo que la probabilidad de acierto que surge de la interacción entre ambos parámetros es de 50% y; por lo tanto, la información de la respuesta sobre la habilidad de la persona está en su máximo valor y el error estándar de la estimación se encuentra en su valor mínimo teórico (Wright y Stone, 1999).

Cuando el valor absoluto de la diferencia entre la habilidad de una persona y la dificultad de un ítem se encuentra en el rango de 1 logit de la habilidad de la persona ($|\theta_n - \beta_i| < 1$), entonces la información de ambos parámetros es mayor a .20. Sin embargo, cuando este valor absoluto de la distancia entre la dificultad del ítem y la habilidad de la persona es mayor a 2 logits off target ($|\theta_n - \beta_i| < 2$), la información es menor que .11 y, si es mayor a 3 logits ($|\theta_n - \beta_i| < 3$), la información será menor a .05 (Wright y Stone, 1979). Esto quiere decir que los ítems con un parámetro de dificultad en la región $|\theta_n - \beta_i| < 1$ aportan más del doble de información para la calibración de ítems y medición de personas que aquellos fuera del rango $|\theta_n - \beta_i| < 2$, y cuatro veces más información que los del rango $|\theta_n - \beta_i| < 3$ (Wright y Stone, 1979).

Los ítems que brindan poca información de la habilidad de la persona se denominan off target. Específicamente, estos ítems pueden interpretarse como aquellos que una persona percibe como muy fáciles o muy difíciles. De esta manera, dichos ítems provocan el surgimiento de conductas no deseables en un proceso de evaluación como la tendencia a la adivinación, falta de interés o mayor ocurrencia de estilos de respuesta (Ingebo, 1997). Dichos comportamientos no deseables afectan la confiabilidad de las puntuaciones obtenidas; por ello, la estrategia targeting es ideal para la mejora de la precisión de la estimación de las medidas (Mallison y Stelmack, 2001).

En la literatura es posible encontrar distintas propuestas en relación a la manera correcta de realizar esta estrategia. Por ejemplo, Boone et al. (2014) consideran que un targeting óptimo se obtiene cuando el promedio aritmético de las dificultades de los ítems y de las habilidades de las personas son similares. Sin embargo, Bond y Fox (2015) sostienen que para que un test se encuentre on target, las medias aritméticas, la distribución y la desviación estándar de los parámetros de dificultad y habilidad deben ser semejantes.

En otra instancia, Linacre (2014; 2019b) afirma que existen dos perspectivas para delimitar ítems on target. Por un lado, en la perspectiva estadística, un ítem se encuentra efectivamente on target cuando la información de la respuesta es mayor o igual a 0.20; tal como se mencionó anteriormente, esto corresponde al rango de $|\theta_n - \beta_i| < 1$ (la diferencia entre habilidad de la persona y dificultad del ítem es menor a 1 logit). En este caso, la probabilidad de acierto se encuentra entre el rango de 28% a 72% o el rango de $\mp 0.94 \approx \mp 1.0$ logits.

Por otro lado, en la perspectiva substancial, Linacre (2014; 2019b) reconoce que utilizar la estrategia estadística ocasiona que los evaluados perciban el test como muy difícil, lo que puede desencadenar conductas no deseables como la adivinación o los estilos de respuesta. En contraste, la perspectiva substancial consiste en situar la media aritmética de los parámetros de dificultad 1 logit por debajo de la media aritmética de habilidad de las

personas. En consecuencia, se obtiene un conjunto de ítems en donde las personas se sienten más cómodas porque tienen un rango de probabilidades de acierto entre 65% y 90%.

Una propuesta alterna fue desarrollada por Wright y Stone (1999); los autores presentaron una categorización del grado en que un test se encuentra on target en función a la diferencia entre las medias aritméticas de los parámetros de habilidad de las personas y dificultad de los ítems. Esta nomenclatura se presenta en la tabla 2.13.

Tabla 2.13

Nomenclatura de targeting según Wright y Stone (1999)

Diferencia entre el promedio de habilidad y dificultad	Categoría
$\lvert\bar{\theta}_n - \bar{\beta}_i\rvert < 1$	On target
$1 < \lvert\bar{\theta}_n - \bar{\beta}_i\rvert < 2$	Lo suficientemente on target
$2 < \lvert\bar{\theta}_n - \bar{\beta}_i\rvert < 3$	Ligeramente off target
$3 < \lvert\bar{\theta}_n - \bar{\beta}_i\rvert < 4$	Off target
$4 < \lvert\bar{\theta}_n - \bar{\beta}_i\rvert$	Extremadamente off target

Nota. Adaptado de *Measurement essentials* (2nd ed., p. 82), por B. Wright y M. Stone, 1999, Wilmington, DE: Wide Range. Copyright 1999 por Benjamin D. Wright y Mark H. Stone.

2.4 La lógica del método Monte Carlo en psicometría

Un experimento de simulaciones busca resolver preguntas complejas a partir de la generación de datos aleatorios que pretenden representar ciertas condiciones de estudio ideales para los objetivos de investigación (Harwell, Stone, Hsu y Kirisci, 1996). Para lograr este objetivo, esta estrategia utiliza el muestreo de datos a partir de múltiples réplicas de experimentos aleatorios (Paxton, Curran, Bollen, Kirby y Chen, 2001).

En breves términos, un experimento aleatorio es un proceso cuyo resultado final no puede ser determinado con certeza; por ejemplo, el lanzamiento de una moneda (Balakrishnan, Koutras y Politis, 2020). Dicha incertidumbre es el motivo por el cual se denomina a muchos fenómenos como variables aleatorias; no obstante, esta nomenclatura no es del todo correcta, pues el fenómeno descrito no es aleatorio ni es variable. De hecho, una variable aleatoria es una función determinista que asigna un número real a cada posible resultado del experimento (también denominado espacio muestral; Grami, 2020).

En otras palabras, si bien no se puede afirmar con certeza el resultado de un evento aleatorio, es posible estimar el posible rango de valores que puede alcanzar (Grami, 2020). Efectivamente, en el campo de la teoría de probabilidades existen diversas aproximaciones para reducir el grado de incertidumbre sobre el resultado del evento (Balakrishnan et al., 2020). Para ello, se utilizan modelos probabilísticos que asocian las variables aleatorias con una distribución de probabilidad para cada posible valor que puede alcanzar (Balakrishnan et al., 2020). De este modo, si se conoce el posible rango de valores que una variable puede alcanzar y la probabilidad de ocurrencia para cada una de ellas, es posible simular dicho experimento (Sobol, 1994).

Luego de establecer un modelo para determinar la distribución de los datos, el siguiente requerimiento para simular un experimento aleatorio es un medio que permita generar números aleatorios [RNG, por sus siglas en inglés]. Ciertamente, los avances en las tecnologías computarizadas han permitido el desarrollo de diversas herramientas para la generación de datos aleatorios (Feinberg y Rubright, 2016). No obstante, todo proceso realizado en una computadora se origina a partir de un algoritmo predefinido; por lo tanto, los datos generados a partir de estas estrategias deberían ser considerados como pseudoaleatorios. En la actualidad existen diversos algoritmos para generar datos que simulan apropiadamente la aleatoriedad; además, estos se encuentra disponibles en diversos softwares de programación libres tales como R (versión 3.5.2; R Core Team, 2018; para

una guía de como simular variables aleatorias en R, consultar Horgan, 2020).

Al contar con un modelo base y un dispositivo generador de número aleatorios, ya es posible simular un experimento aleatorio; por ejemplo, una distribución normal estándar es un modelo en donde la función de probabilidad se encuentra determinada por un parámetro de localización, representado por la media aritmética con un valor convencional de 0, y un parámetro de forma, representado por la varianza (o en algunos casos como la desviación estándar) con un valor de 1 (Grami, 2020).

Una de las propiedades de esta distribución es la simetría sobre la media aritmética, es decir que ambos extremos de la distribución son iguales. De este modo, es posible afirmar que el 50% de los datos de la distribución se encuentran por encima y por debajo de la media aritmética (ver figura 2.9).

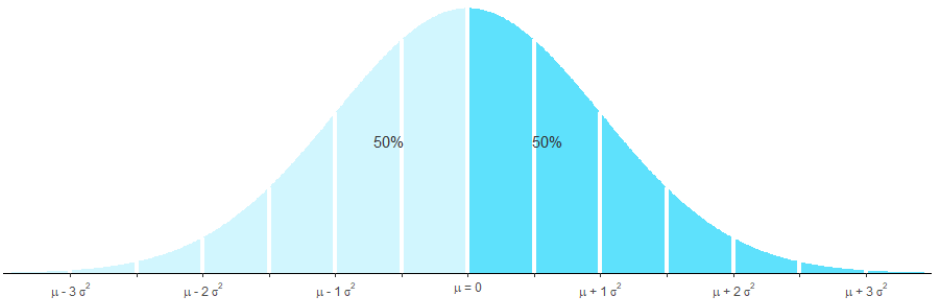

Figura 2.9. La propiedad de simetría de la Distribución Normal Estándar.

A través del método Monte Carlo es posible contrastar esta propiedad al muestrear un número determinado de casos a partir de dicha distribución y estimar la proporción de valores menores (o mayores) a 0. Si dicha propiedad se cumple, entonces el resultado de este procedimiento debería ser .50, que indica el 50% de los datos. Este experimento es fácilmente replicable en el software R (versión 3.5.2; R Core Team, 2018), utilizando las siguientes líneas de código:

```
n<- 1000 #Número de casos a simular

a<- rnorm(n,0,1) #Función que simula datos de una distribución normal

sum(a<0)/n #Proporción de valores menores a 0.
```

En otro ejemplo desarrollado utilizando el modelo Rasch, al considerar el caso en que se pretende simular la respuesta que podría generar una persona con una habilidad de 1.5 a un ítem dicotómico con dificultad de 1.0, el procedimiento inicial sería determinar la probabilidad de acierto de dicha persona. De acuerdo al modelo matemático, determinar la probabilidad es relativamente simple una vez que se conocen los parámetros involucrados, pues solo es necesario reemplazar los valores en la ecuación.

$$P_{ni}(x_{ni} = 1|\theta_n, \beta_i) = \frac{e^{(\theta_n = 1.5 - \beta_i = 1.0)}}{1 + e^{(\theta_n = 1.5 - \beta_i = 1.0)}}$$

$$P_{ni}(x_{ni} = 1|\theta_n, \beta_i) = \frac{e^{(0.5)}}{1 + e^{(0.5)}}$$

$$P_{ni}(x_{ni} = 1|\theta_n, \beta_i) = \frac{1.65}{1 + 1.65}$$

$$P_{ni}(x_{ni} = 1|\theta_n, \beta_i) = 0.62$$

Ahora, como solo existen dos posibles resultados para la respuesta a un ítem dicotómico (acierto o fallo), entonces es posible simular el experimento aleatorio como un ensayo de Bernoulli con un parámetro $p = 0.62$ (o un experimento Binomial con $p = .62$ y $n = 1$). Aunque esta aproximación sea relativamente simple, es la noción básica que subyace a los estudios Monte Carlo en psicometría.

En efecto, dada la naturaleza probabilística de los modelos TRI y Rasch, es posible realizar múltiples experimentos aleatorios a partir del control de los parámetros del modelo (Feinberg y Rubright, 2016). La utilidad de esta

estrategia puede observarse en distintos ámbitos; por ejemplo, al evaluar los distintos métodos de estimación de parámetros. En la práctica es imposible conocer el parámetro verdadero de una persona o ítem; por lo tanto, determinar qué método de estimación es el más preciso parecería un objetivo inalcanzable.

Sin embargo, al utilizar simulaciones Monte Carlo, los parámetros de las personas e ítems pueden ser creados artificialmente, tal como se realizó en el ejemplo anterior. A partir de dichos parámetros es posible generar múltiples conjuntos de datos de respuestas a ítems a través del experimento binomial previamente descrito. En consecuencia, si cada conjunto de datos simulados se somete a los métodos de estimación que se pretenden comparar, aquel método que provea estimaciones más cercanas a los parámetros verdaderos puede ser considerado como el más preciso (e.g., Wang y Wang, 2002).

Efectivamente, la comparación de la precisión de distintos métodos de estimación es uno de los estudios más comunes en psicometría (Harwell et al., 1999). Adicionalmente, también se aplica dicho procedimiento para determinar el tamaño de muestra más idóneo para implementar un análisis estadístico, comparar distintas estrategias de equiparación o detección de funcionamiento diferencial de los ítems, evaluar los efectos del incumplimiento de supuestos de dimensionalidad en diversas técnicas estadísticas, entre otros (Bulut y Sünbül, 2017; Harwell et al., 1996; Feinberg y Rubright, 2016; Paxton et al., 2001).

En cuanto a la metodología experimental que subyace a un experimento Monte Carlo, existen diversas propuestas en la literatura (e.g., Bulut y Sünbül, 2017; Harwell et al., 1996; Feinberg y Rubright, 2016; Paxton et al., 2001); sin embargo, el planteamiento e implementación de dicho experimento dependerá de los objetivos de la investigación y del ámbito en el que se circunscribe. Un mayor detalle sobre la metodología empleada en este estudio se especifica en el capítulo de metodología.

CAPÍTULO III: OBJETIVOS, HIPÓTESIS Y DEFINICIÓN DE VARIABLES

3.1 Objetivo

Comparar la confiabilidad obtenida a partir de la aplicación de un test on target para toda una población con aquella obtenida al utilizar tests on target para dos grupos poblacionales en situaciones que varían de acuerdo a la distancia entre las medias aritméticas de los parámetros de habilidad de ambos grupos.

3.2 Hipótesis

La confiabilidad de las puntuaciones obtenidas a partir de tests on target para dos grupos poblacionales será mayor a la obtenida a partir de un solo test on target para toda la población. Además, conforme la distancia entre las medias aritméticas de habilidad de ambos grupos incremente, mayor será la ganancia en confiabilidad.

3.3 Definición de variables

En el presente estudio se propone un diseño experimental factorial en donde se incluye como variable dependiente a la confiabilidad de las puntuaciones y como variables independientes al target de la evaluación y a la diferencia entre medias aritméticas de parámetros de habilidad de los grupos que componen la población.

3.3.1 Confiabilidad

La confiabilidad se refiere a la consistencia y precisión de las puntuaciones obtenidas a través de múltiples réplicas del proceso de evaluación (AERA et al., 2014; Coolican, 2014; Desjardins y Bulut, 2018; Urbina, 2014). Además, esta noción es considerada como uno de los principios fundamentales de la medición a través de pruebas psicométricas (Geisinger, 2013); y es empleada como criterio para la evaluación de la calidad de instrumentos de medición psicológicos (Cohen y Swerdlik, 2010).

Para su operacionalización, se utilizó el índice de confiabilidad de separación de personas, un coeficiente del modelo de medición Rasch (Andrich y Marais, 2019). Este índice indica el grado en que es posible replicar el ordenamiento de las personas en el continuum latente si se reevalúa a las personas a través de otro conjunto de ítems (Bond y Fox, 2015). Matemáticamente, este coeficiente se expresa como la fracción de la varianza observada de las respuestas que puede ser reproducida por el modelo. Un valor menor a .50 sugiere que las diferencias entre las medidas pueden explicarse principalmente por el error de medición (Fisher, 1992).

3.3.2 Target

Los instrumentos psicométricos se componen a partir de un conjunto de ítems que reflejan muestras del comportamiento y permiten la medición indirecta de fenómenos no observables (Cohen y Swerdlick, 2010). Una de las estrategias de ensamblaje de ítems para formar pruebas es el targeting, que consiste en seleccionar los ítems más relevantes para el nivel de habilidad de la persona o grupo objetivo (Bond y Fox, 2015; Wright y Stone, 1979).

En la literatura se presentan diversas propuestas para implementar un targeting apropiado. En este experimento se utilizó el criterio sugerido por Bond y Fox (2015) que consiste en que el rango de parámetros de dificultad de los ítems debe aproximarse al rango de los parámetros de habilidad de las personas. Como la simulación de parámetros de habilidad de las

personas se realizó considerando una distribución normal estándar, la cal es la distribución más utilizada para simular parámetros de habilidad en estudios Monte Carlo en psicometría (Harwell et al., 1996), el rango apropiado de dificultad idóneo para los ítems es de -2.5 a 2.5 desviaciones estándar alrededor de la media aritmética de los parámetros de habilidad. El fundamento que subyace a la estrategia targeting es que, en una distribución normal, más del 95% de los datos se encuentra en este rango de valores, específicamente 98.8% (Andrich y Marais, 2019; Wilcox, 2010).

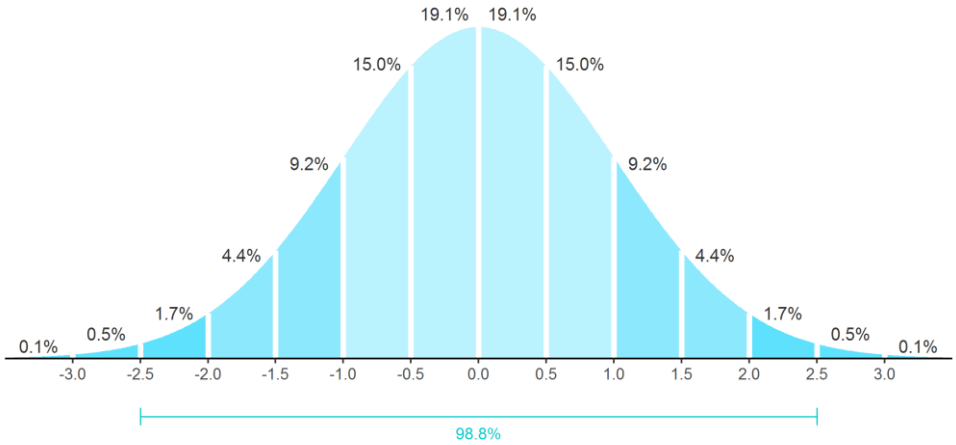

Figura 3.1. La distribución normal estándar. La línea inferior indica el rango de valores establecido para los parámetros de ítems del experimento, que representan el 98.8% de la totalidad de los parámetros simulados de las personas.

Para reafirmar la idoneidad de la estrategia de targeting designada, se solicitó la opinión de John M. Linacre, uno de los expertos más reconocidos en el marco del modelo de medición Rasch y creador del software *WINSTEPS*. Linacre administra personalmente el portal web *raschforum.boards.net*, en donde responde dudas sobre los procesos de análisis y medición relacionados al el modelo Rasch. A través de este medio, Linacre (2019b) aprobó el criterio designado y sugirió ciertas recomendaciones para el experimento.

3.3.3 Distancia entre medias aritméticas del nivel de habilidad

La habilidad representa el nivel o ubicación de una persona en un continuum del rasgo latente evaluado (Kline, 2015). En este estudio, la distribución de habilidad de ambos grupos poblacionales se simuló a partir de una distribución normal con una misma desviación estándar $\sigma^2 = 1$ y una media aritmética que sería manipulada para establecer cada condición de la simulación. Para este factor se designaron cuatro instancias de diferencias entre las medias aritméticas de los grupos que componen la población: 0.5, 1.0, 1.5 y 2.0 desviaciones estándar. Como apoyo gráfico, los niveles de esta variable se presentan en las figuras 3.2 a 3.5.

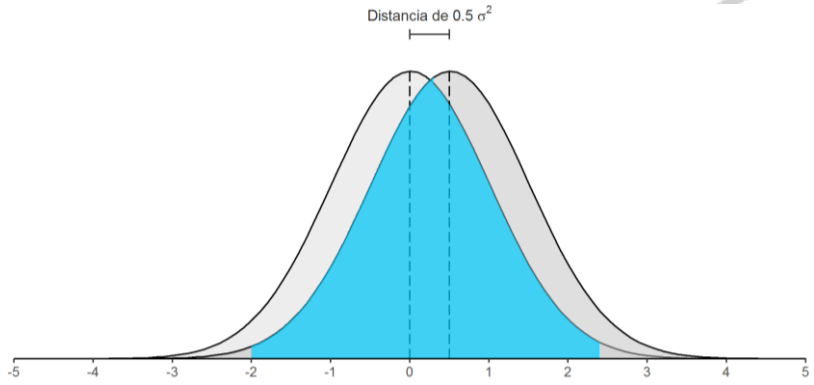

Figura 3.2. El primer nivel de la variable independiente diferencia entre medias aritméticas del nivel de habilidad de dos grupos.

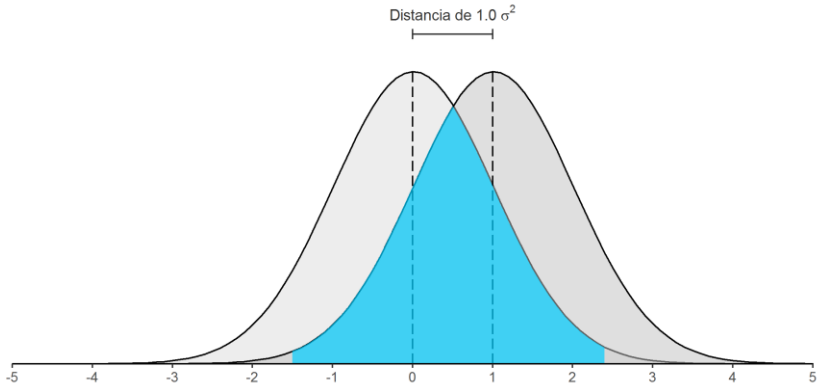

Figura 3.3. El segundo nivel de la variable independiente diferencia entre medias aritméticas del nivel de habilidad de dos grupos.

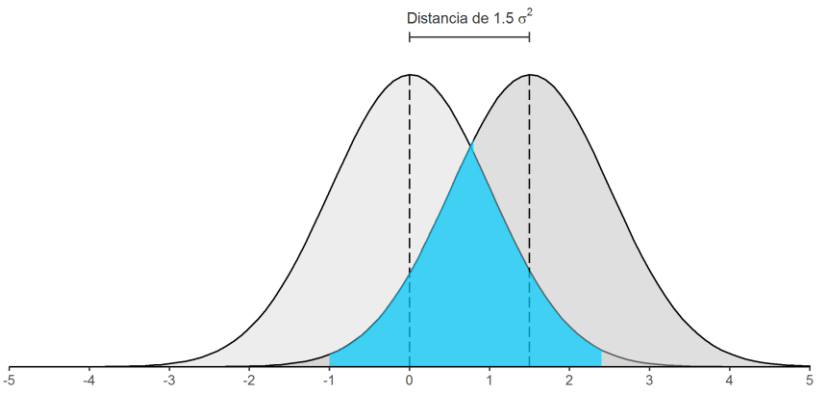

Figura 3.4. El tercer nivel de la variable independiente diferencia entre medias aritméticas del nivel de habilidad de dos grupos.

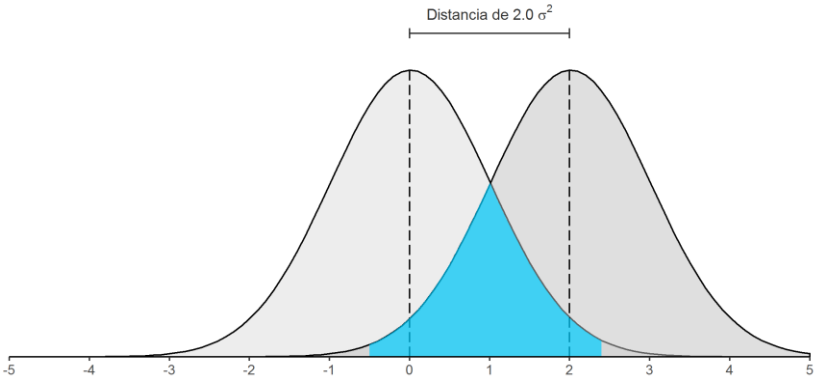

Figura 3.5. El cuarto nivel de la variable independiente diferencia entre medias aritméticas del nivel de habilidad de dos grupos.

CAPÍTULO IV: MÉTODO

En la presente investigación se utilizó el método de simulaciones Monte Carlo con el fin del lograr el objetivo de comparar la confiabilidad obtenida a partir de la aplicación de un test on target para toda una población con aquella obtenida al utilizar tests on target para dos grupos poblacionales en situaciones que varían de acuerdo a la distancia entre las medias aritméticas de los parámetros de habilidad de ambos grupos.

Fundamentalmente, un experimento de simulaciones busca resolver preguntas complejas a partir de la generación de datos aleatorios que pretenden representar ciertas condiciones de estudio ideales (Harwell et al., 1996). Sin embargo, es importante mencionar que no existe una secuencia estandarizada acerca del método Monte Carlo, pues su aplicación varía en relación al objetivo del estudio y a la disciplina en donde se utilice la técnica (Harrison, 2010; Olvera, 2017).

En particular, el procedimiento delimitado para implementar el estudio experimental de simulaciones se planteó sobre la base de las recomendaciones presentadas en los trabajos de Bulut y Sünbül (2017), Harwell et al. (1996) y Feinberg y Rubright (2016). Estas publicaciones presentan una serie de directrices para la conducción apropiada de experimentos Monte Carlo en estudios psicométricos. Debido a la relativa complejidad de estos los estudios de simulación, en la figura 4.1 se presenta un resumen de las fases propuestas para el experimento como guía para comprender la metodología designada.

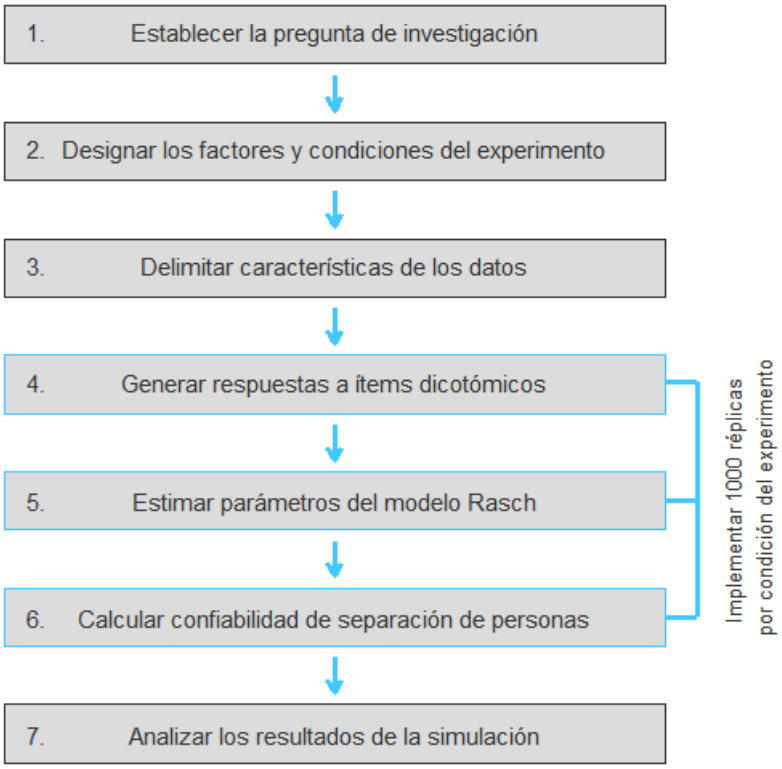

Figura 4.1. Las fases del experimento Monte Carlo propuesto. Como ya se mencionó, no existe una estructura estandarizada de las fases de un experimento de simulaciones, esta propuesta se realiza sobre la base de los trabajos de Bulut y Sünbül (2017), Harwell et al. (1996) y Feinberg y Rubright (2016).

La primera fase al implementar un experimento Monte Carlo es la misma que en cualquier otro tipo de investigación; consiste en el planteamiento de una *incógnita* que se desea resolver (Harwell et al., 1996). Particularmente, esta fase se delimitó en el primer capítulo de este reporte de investigación, en general, el objetivo de este estudio es determinar el efecto de la diferencia entre las medias aritméticas de la distribución de habilidad de grupos poblacionales en la confiabilidad de las puntuaciones.

El siguiente paso consiste en delimitar los *factores* y *condiciones* de la simulación. En el contexto de los experimentos Monte Carlo, los factores son similares a las variables independientes en un diseño experimental y su interacción delimita las distintas condiciones del experimento (Feinberg y Rubright (2016). Las variables dependientes fueron especificadas en el tercer capítulo de este documento; sin embargo,

un detalle más específico acerca del diseño experimental empleado se presenta en la sección de *tipo y diseño de investigación*.

La tercera fase implica establecer ciertas especificaciones acerca de la naturaleza de los datos que se desean simular en un experimento Monte Carlo. En este estudio, el modelo Rasch para ítems dicotómicos fue designado como la base para la generación de datos aleatorios (i.e., respuestas a ítems dicotómicos). Adicionalmente, el modelo involucra la interacción entre parámetros de habilidad y dificultad, cuyas características distribucionales se especifican en las secciones *participantes* y *técnicas de recolección de datos*, respectivamente.

La tercera fase implica establecer ciertas especificaciones acerca de la naturaleza de los datos que se desean simular en un experimento Monte Carlo. En este estudio, el modelo Rasch para ítems dicotómicos fue designado como la base para la generación de datos aleatorios (i.e., respuestas a ítems dicotómicos). Adicionalmente, el modelo involucra la interacción entre parámetros de habilidad y dificultad, cuyas características distribucionales se especifican en las secciones *participantes* y *técnicas de recolección de datos*, respectivamente.

Al culminar estas primeras fases, ya es posible implementar la generación de datos aleatorios, la estimación de parámetros y la estimación de la confiabilidad en cada condición del experimento. En particular, estas tres fases son replicadas un número determinado de veces (i.e., 1000). Todas las especificaciones con respecto a este procedimiento iterativo se detallan en la sección de *procedimiento para la recolección de datos*.

Finalmente, todo el experimento culmina en el análisis de los datos obtenidos en cada condición del experimento sobre la base de los objetivos de la investigación. En esta última fase, se emplean los análisis descriptivos e inferenciales necesarios para comparar la confiabilidad obtenida en pruebas on target para dos grupos poblacionales con aquella obtenida a partir de una prueba única on target para toda la población; los resultados se presentan en el capítulo V. Tanto la generación de datos como los análisis respectivos fueron realizados en el lenguaje de programación R (versión 3.5.2; R Core Team, 2018).

4.1 Tipo y diseño de investigación

La presente investigación busca comparar la confiabilidad obtenida al utilizar tests on target para dos grupos poblacionales con aquella obtenida a partir de la aplicación de un test on target para toda la población en situaciones que varían de acuerdo a la distancia entre las medias aritméticas de los parámetros de habilidad de dichos grupos. En otras palabras, se busca analizar el efecto que ejerce la distancia entre las medias aritméticas de ambas distribuciones de habilidad en la confiabilidad de las puntuaciones. Cuando el objetivo de una investigación es estudiar relaciones causales entre variables, esta se denomina investigación experimental *verdadera* (Bordens y Abbott, 2018; Creswell y Creswell, 2018; Edmonds y Kennedy, 2017; Johnson y Christensen, 2016; Salkind, 2018).

La característica fundamental de las investigaciones experimentales es el alto grado de control que se requiere sobre los factores que podrían influir en el objeto de estudio (Creswell y Creswell, 2018). Una alternativa que permite ejercer dicho grado de control sobre distintos factores y estudiar su efecto de manera simultánea, son los estudios de simulaciones Monte Carlo (Harwell et al., 1996). En efecto, los experimentos de simulaciones utilizan el muestreo aleatorio de datos para analizar diversas condiciones establecidas a partir de la interacción entre los factores considerados en el estudio (Feinberg y Rubright, 2016).

Con el fin de asegurar la idoneidad de los experimentos Monte Carlo, Harwell et al. (1996) consideran que dichos estudios deben desarrollarse sobre la base de un diseño apropiado que sea congruente con los objetivos de la investigación. Entre diversas alternativas, los autores sugieren utilizar un diseño factorial cuando se consideran pocas variables independientes en el estudio. En este sentido, como solo se incluyen las variables independientes *target* de la evaluación y *diferencia entre medias aritméticas del nivel de habilidad*, para el estudio de la variable dependiente *confiabilidad*, un diseño factorial representa una alternativa apropiada.

A continuación, se presenta el diseño factorial que subyace al experimento Monte Carlo planteado en esta investigación. Para ello, se utilizó la terminología desarrollada por Rex B. Kline (2009):

$$R \quad X_{A1B1} \quad O_1$$

$$R \quad X_{A2B1} \quad O_2$$

$$R \quad X_{A1B2} \quad O_3$$

$$R \quad X_{A2B2} \quad O_4$$

$$R \quad X_{A1B3} \quad O_5$$

$$R \quad X_{A2B3} \quad O_6$$

$$R \quad X_{A1B4} \quad O_7$$

$$R \quad X_{A2B4} \quad O_8$$

En donde R se refiere a un grupo que resulta de una asignación aleatoria establecida en el experimento, cada grupo representa a una población de 5000 casos simulados. X representa la exposición de dicho grupo a determinadas condiciones que se constituyen a partir de la interacción entre los distintos niveles de las variables independientes.

La variable target (representada por A) se compone por dos niveles, en donde A_1 representa a una evaluación compuesta por dos pruebas on target para dos grupos que componen la población; mientras que, A_2 representa a una evaluación en donde solo se emplea una prueba on target para toda la población.

La variable diferencia entre medias aritméticas de los niveles de habilidad de dos grupos que componen la población (representada por B) se constituye a partir de cuatro niveles: B_1, B_2, B_3 y B_4. Cada uno de ellos representa una distancia entre medias aritméticas de habilidad; específicamente, 0.5, 1.0, 1.5 y 2.0 desviaciones estándar, respectivamente.

Finalmente, O representa a la observación de la variable dependiente confiabilidad en cada interacción entre las variables independientes. En otras palabras, al establecer un diseño 2x4, la interacción entre las variables independientes delimita un total de ocho condiciones para el experimento, en donde cada una tendrá una medida de confiabilidad, esto se resume en la tabla 4.1.

Tabla 4.1

Factores y condiciones empleados en el estudio de simulación

Distancia entre medias \ Tipo de prueba	Enfocada (targeted) en los grupos	Enfocada (targeted) en la población
0.5 desviaciones estándar	O_1	O_2
1.0 desviaciones estándar	O_3	O_4
1.5 desviaciones estándar	O_5	O_6
2.0 desviaciones estándar	O_7	O_8

4.2 Participantes

En esta sección se describen todos los criterios considerados con el objetivo de establecer un tamaño de muestra idóneo que delimite el número de casos a simular en cada condición del experimento Monte Carlo. Posteriormente, se presenta el modelo distribucional a partir del cual se generaron parámetros de habilidad para cada caso simulado, tomando en cuenta las los distintos niveles de las variables independientes establecidas en el experimento.

Para determinar el número de casos en la simulación, se consideraron distintos aspectos. Primero, se tomó en cuenta la cantidad mínima requerida para obtener un tamaño del efecto de 0.5, que representa al valor más bajo de la distancia de medias propuesta en este experimento. Segundo, como en el presente estudio se emplea el modelo de medición Rasch, se consideraron los tamaños de muestra necesarios para la estabilidad y precisión de los parámetros de la dificultad de ítems y habilidad de personas. Finalmente, como criterio más relevante para la delimitación del tamaño de muestra, se utilizó el principio de autenticidad del método Monte Carlo, el cual sostiene que las condiciones del experimento deben representar lo mejor posible la realidad que se pretende simular.

En primer lugar, se utilizó el programa G*Power (versión 3.1.9.2) de Faul, Erdfelder, Lang y Buchner (2009) para delimitar un tamaño de muestra apropiado para la comparación de dos muestras independientes. En este caso se tomó en cuenta el tamaño del efecto de 0.5, el cual es el mínimo valor utilizado en este experimento y un nivel de significancia de .05. Además, se consideró una potencia estadística de .95, que es mayor al mínimo valor aceptable de .80 propuesto por Cohen (1992) para la investigación en las ciencias del comportamiento; el resultado final fue de 176 casos por grupo.

En segundo lugar, se revisaron diversos estudios acerca de la precisión y estabilidad de la estimación de parámetros. Uno de ellos fue el trabajo de Linacre (1994) *Sample size and item calibration [or person measure] stability* en donde delimita la importancia del tamaño de muestra para la

estabilidad del cálculo de los parámetros del modelo. El autor enfatiza que cada vez que se calibra un conjunto de ítems en muestras distintas de personas con características similares, se suelen esperar resultados distintos. Cuando las muestras son muy pequeñas, los resultados pueden ser muy inestables, ya que el modelo Rasch es similar a otros análisis estadísticos cuando se utilizan en muestras insuficientes, se producen estimaciones poco precisas y menos robustas.

Linacre (1994) afirma que el tamaño de muestra mínimo aceptable para la estimación estable de parámetros en tests compuestos por ítems dicotómicos es 30 casos, obteniendo una estabilidad de ±1 logits. Adicionalmente, es importante reconocer que conforme aumente el número del tamaño de la muestra, mayor será la estabilidad de la calibración de ítems o medidas de personas, tal como se muestra en la tabla 4.2.

Tabla 4.2

Tamaño de muestra mínimo y estabilidad de los parámetros

Estabilidad de la calibración de ítems y medidas de las personas.	Intervalo de confianza	Rango de tamaño de muestra mínimo.
± 1 logit	95%	16-36
± 1 logit	99%	27-61
± 0.5 logit	95%	64-144
± 0.5 logit	99%	250—20*longitud del test

Nota. Adaptado de "Sample size and item calibration [or person measure] stability", por J. M. Linacre, 1994, *Rasch Measurement Transactions, 7*(4), p. 328. Copyright 1994 por John Michael Linacre.

Además, diversos autores han afirmado que el modelo Rasch permite estimar parámetros con utilidad práctica incluso en tamaños de muestra pequeños (Linacre, 1996; 2005; Lord, 1983; Wright, 1977; Wright y Douglas, 1975). Estas declaraciones han sido respaldadas por un amplio

número de estudios experimentales que utilizaron el método Monte Carlo para determinar el efecto del tamaño de la muestra en la estimación de parámetros del modelo encontraron que tamaños de muestra relativamente pequeños $n < 500$ son suficientes para una estimación útil (Baur y Lukes, 2009; Custer, 2015; Goldman y Raju, 1986; Guyer y Thompson, 2011; Sahin y Aml, 2016; Stone y Yumoto, 2004; Svetina et al., 2013).

En congruencia con estos estudios, Linacre (1994) considera que cuando se cuenta con una muestra extensa de 2000 o 3000 evaluados, el cálculo de los parámetros sería esencialmente idéntico en ambas muestras. El problema es que en la práctica resulta muy costoso para los investigadores alcanzar un tamaño de muestra tan amplio. En el presente estudio, designar un tamaño de muestra extenso no representa un problema substancial porque el método Monte Carlo permite la simulación de la cantidad de casos que desee el investigador, considerando que un amplio tamaño de muestra puede complejizar los cálculos.

El tercer criterio para delimitar el tamaño de la muestra fue el principio de *autenticidad* del método Monte Carlo, que se refiere al grado en que el experimento propuesto refleja las condiciones de la realidad que se pretende simular (Bulut y Sünbül, 2017). La investigación tiene como objetivo simular poblaciones extensas compuestas por grupos relevante, especialmente un proceso de evaluación educativa a gran escala, en donde se suelen emplear tamaños de muestra extensos (Cresswell et al., 2015).

Con el objetivo de representar en mayor grado a las evaluaciones educativas a gran escala, se consideraron las características de distintos programas de evaluación a nivel internacional. Uno de ellos fue el *Programa para la Evaluación Internacional de Alumnos* [PISA, por sus siglas en inglés] de la Organization for Economic Co-operation and Development [OECD]. Para asegurar la representatividad a nivel de país, PISA 2018 consideró una muestra mínima de 6300 estudiantes para formar parte de la evaluación computarizada y 5250 para una evaluación en formato de lápiz y papel (OECD, 2016).

En otra instancia, las evaluaciones del *Progress in International Reading Literacy Study* [PIRLS] requieren de un tamaño de muestra mínimo de 4000 estudiantes por cada grado objetivo en la mayoría de países (LaRoche, Joncas y Foy, 2016). Del mismo modo, el programa de evaluación internacional *Trends in International Mathematics and Science Study* [TIMSS] también establece un tamaño de muestra mínimo de 4000 estudiantes por cada grado objetivo, aunque este valor puede ser mayor dependiendo del número promedio de estudiantes por salón del país (Joncas y Foy, 2012).

También se consideró el tamaño de muestra delimitado en el diseño del *Tercer Estudio Regional Comparativo y Explicativo* [TERCE] del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [LLECE]. En general, la muestra de estudiantes peruanos osciló entre 4739 y 5038, tal como se presenta en la tabla 4.3.

Tabla 4.3

Muestra efectiva de estudiantes peruanos según área evaluada

Área evaluada	Tamaño de muestra
Lectura (3er grado)	4946
Lectura (6to grado)	4739
Matemática (3er grado)	5038
Matemática (6to grado)	4789
Ciencia (6to grado)	4801
Escritura (3er grado)	5003
Escritura (6to grado)	4745

Nota. Adaptado de "Informe de resultados TERCE: Logros de aprendizaje," por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [LLECE], 2015 (http://unesdoc.unesco.org/images/0024/002435/243532S.pdf). Copyright 2016 por UNESCO.

Tomando en cuenta los tres criterios mencionados, se optó por delimitar un tamaño de muestra de 5000 casos para la simulación de parámetros de habilidad de personas para la población total y; al mismo tiempo, se designó que esta población se encontraría compuesta por dos grupos relevantes de 2500 casos en cada uno. La equidad de casos en cada grupo poblacional se estableció con el fin de controlar el posible efecto del tamaño de muestra en la precisión de la estimación de parámetros y; consecuentemente, de la confiabilidad de las puntuaciones.

Con respecto al modelo distribucional de los parámetros de habilidad de la población, se estableció una distribución normal estándar para cada grupo que compone la población. Este es el modelo distribucional más utilizado para el parámetro de habilidad θ de una población en estudios Monte Carlo desarrollados en el contexto de modelos TRI (Harwell et al., 1996). En general, la distribución normal requiere de un parámetro de localización μ (media aritmética) y otro de forma σ^2 (varianza; Grami, 2020). Para ambos grupos, se estableció el mismo parámetro de forma ($\sigma^2 = 1$) y diferentes parámetros de localización, presentados en la tabla 4.4.

Tabla 4.4

Medias aritméticas para los grupos poblacionales

Diferencia entre medias aritméticas	Media aritmética en el grupo A	Media aritmética en el grupo B
0.5	0.0	0.5
1.0	0.0	1.0
1.5	0.0	1.5
2.0	0.0	2.0

.

Para simular los parámetros de habilidad para los dos grupos que componen la población, se utilizó la función *rnorm*() en R (versión 3.5.2; R Core Team, 2018). Esta herramienta permite generar datos aleatorios sobre la base de una distribución normal, considerando la delimitación previa de una media aritmética y una desviación estándar como parámetros del modelo distribucional para cada grupo que compone la población.

Finalmente, la distribución de habilidad de toda la población se realizará a partir de la simple concatenación de los parámetros de habilidad de ambos grupos, de modo que el rango de habilidad sea el mismo para ambos niveles de la variable independiente target. Como Linacre (2019a) afirma, la variabilidad de los rangos de habilidad puede afectar al cálculo del índice de confiabilidad de separación de personas y; por lo tanto, no controlar esta variable puede sesgar a los resultados del estudio.

4.3 Técnicas de recolección de datos

En las evaluaciones psicométricas empíricas, la técnica de recolección de datos por defecto suele ser un test psicométrico. Como analogía ante esta tendencia, en esta sección se describen las características de los tests on target diseñados para cada condición del experimento Monte Carlo, con especial énfasis en el número de ítems y la distribución de los parámetros de dificultad.

Claramente, existen distintas posturas en relación a la extensión más adecuada de un test. Algunos autores argumentan que, conforme aumente el número de ítems en un test, mayor será la confiabilidad de las puntuaciones obtenidas (Urbina, 2014). En consonancia con esta afirmación, Kruyen (2012) sostiene que utilizar tests con pocos ítems puede sesgar los procesos de toma de decisiones sobre individuos; por ello, el autor propone ensamblar tests a partir de un mínimo de 20 ítems.

En contraste, Wright (1992) considera que no es apropiado delimitar un número mínimo de ítems para asegurar la confiabilidad de las puntuaciones. En su lugar, el autor considera que un desarrollador de un test debería considerar la cantidad necesaria de ítems que permita realizar inferencias en relación a los usos propuestos para los que se diseña y administra un test; en otras palabras, una cantidad de ítems que represente apropiadamente el contenido que se pretende medir (AERA et al., 2014). Efectivamente, Lunz (2009) demostró que una mayor cantidad de ítems no siempre implica una mejora en la precisión de las puntuaciones, pues la confiabilidad depende

de otras variables además de la longitud, como la calidad de los ítems que componen la escala.

Debido a la controversia presente en la literatura en relación al número óptimo de ítems para un test; en este experimento se consideró una revisión de la literatura empírica sobre la longitud del test apropiada para los modelos Rasch (o TRI de un parámetro) y; nuevamente, el principio de autenticidad del método de simulación Monte Carlo como guía para delimitar un número idóneo de ítems.

En general, no parece haber un consenso claro en relación al número idóneo de ítems para un test en los resultados de los estudios de simulaciones. Diversos experimentos en donde se ha estudiado el efecto del número de ítems de un test en la precisión de la estimación de parámetros concluyen que existe un efecto considerable de dicho factor (Baur y Lukes, 2009; Khan, 2014; Lord, 1983). No obstante, las recomendaciones establecidas sobre la base de los resultados de estos estudios oscilan en un rango entre 8 (Svetina et al., 2013) a 50 ítems (Guyer y Thompson, 2011) como mínimo.

En vista de esta discusión, se procedió a analizar las tendencias sobre el número de ítems establecido en las evaluaciones educativas a gran escala, en la tabla 4.5 se presentan algunas aproximaciones nacionales e internacionales.

Tabla 4.5

Número de ítems evaluados según dominio en diversas evaluaciones educativas a gran escala a nivel nacional e internacional

Evaluación	Dominio	Número de ítems
Programa para la Evaluación Internacional de Alumnos [PISA] (2015). Jóvenes de 15 años.	Matemática	69
	Lectura	88
	Ciencias	85
Segundo Estudio Regional Comparativo y Explicativo [SERCE] (2006). Sexto grado de primaria.	Matemática	87
	Lectura	96
	Ciencias	84
Tercer Estudio Regional Comparativo y Explicativo [TERCE] (2013). Sexto grado de primaria.	Lectura	94
	Matemática	90
	Ciencias	82
Evaluación Censal Estudiantil [ECE] (2015). Segundo de secundaria	Matemática	90
	Lectura	70
Evaluación Censal Estudiantil [ECE] (2016). Segundo grado de secundaria	Matemática	90
	Lectura	86
	Historia, Geografía y Economía	86

Nota. Los datos fueron obtenidos de los reportes técnicos de cada evaluación y pueden ser consultados en: OECD (2017); UNESCO y Oficina Regional de Educación para América Latina y el Caribe [OREALC] (2009; 2016); y MINEDU (2018).

Es importante mencionar que los estudiantes no se enfrentan a la totalidad de los ítems en estos programas de evaluación estandarizada, ya que la mayoría de ellos emplea procesos complejos de muestreo de matriz para generar formas distintas del test. En otras palabras, cada estudiante se enfrenta a un subconjunto de ítems que luego es transformado a la misma métrica por métodos de equiparación (Gonzales y Rutkowski, 2010).

Tomando en cuenta ambos criterios, se optó por utilizar un total de 40 ítems para la simulación de todos los tests on target del experimento, esto incluye a las dos pruebas que componen la evaluación a partir de tests on target enfocados en los grupos de la población y a la prueba única que representa la evaluación a partir de un tests on target enfocado en toda la población. El

motivo por el cual se tomó la decisión de delimitar el mismo número de ítems para la prueba de ambos grupos y la de toda la población es porque Linacre (2019a) afirma que la estimación del índice de confiabilidad de separación de personas puede verse afectado por la extensión del test; por ello, mantener el mismo número de ítems para cada evaluación permite controlar el efecto de esta variable.

La estrategia empleada para realizar el targeting fue una adaptación de la propuesta de Bond y Fox (2015), la cual establece que el rango de parámetros de dificultad debe aproximarse al rango de los parámetros de habilidad de las personas. Como la simulación de parámetros de habilidad se realiza sobre la base de una distribución normal, el rango apropiado de dificultad escogido es de -2.5 a 2.5 desviaciones estándar alrededor de la media de habilidad, esto se realiza porque en una distribución normal, más del 95% de los datos se encuentra en este rango (Andrich y Marais, 2019; Wilcox, 2010).

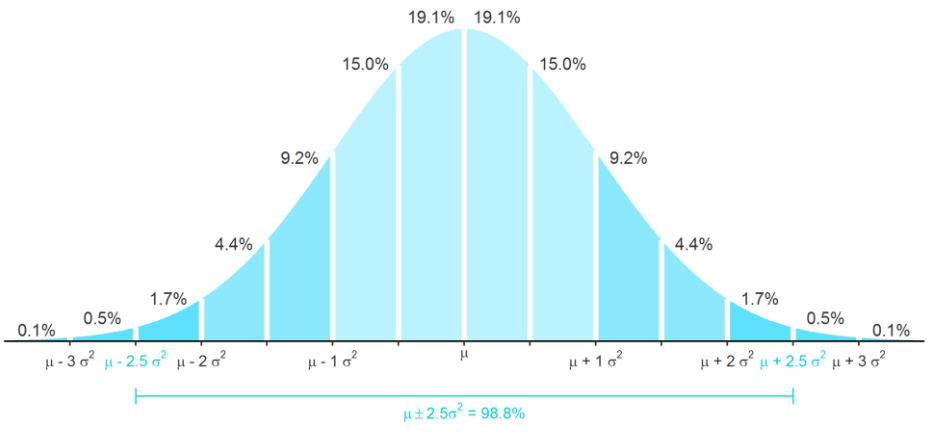

Figura 4.2. El rango de parámetros de dificultad para un targeting apropiado para la distribución normal de la variable latente. Como se mencionó anteriormente, este método fue aprobado por John M. Linacre (2019b).

Una cuestión importante antes de delimitar la distribución de los parámetros de dificultad de los ítems es que los tests on target para grupos poblacionales implica considerar que las pruebas on target para los grupos de la población deben encontrarse en la misma métrica; por ello, es indispensable realizar un proceso de equiparación de puntuaciones. En el contexto de los modelos Rasch o TRI, dos formas del test pueden ser fácilmente equiparadas si cuentan con un conjunto de ítems en común (Boone, 2016; Kolen y Brenan, 2014). A esta estrategia se le conoce como *calibración concurrente* y consiste en estimar los parámetros de los ítems de ambas formas en conjunto; para ello, los ítems únicos son tratados como ítems no alcanzados por los grupos que no se enfrentaron a ellos (Gonzales y Wiberg, 2017).

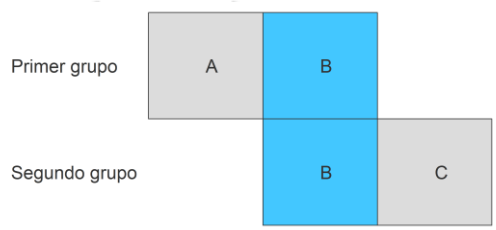

Figura 4.3. Diseño de bloques del experimento. El bloque de ítems B representa a los ítems en común a los que ambos grupos se enfrentarán. Además, el primer grupo se enfrentará al bloque de ítems A; mientras que los ítems del bloque C se consideran como no alcanzados para este grupo. Del mismo modo, el segundo grupo se enfrentará al bloque de ítems C; mientras que los ítems del bloque A serán considerados como no alcanzados para este grupo.

En la figura 4.3 se representa gráficamente el diseño de bloques empleado para este experimento. Al tener una estructura clara para asegurar la equiparación de las medidas, el siguiente paso fue delimitar los parámetros de dificultad para todos los ítems. Para lograr este objetivo, primero se designó un rango del continuum latente que correspondería a las dificultades de los ítems en común del bloque B y otro para las dificultades de los ítems únicos del bloque A para el primer grupo, y del bloque C para el segundo grupo.

Como el criterio utilizado para el targeting consiste en delimitar un rango de -2.5 y 2.5 desviaciones estándar alrededor de la media aritmética de la distribución de habilidad, un sector de este rango será destinado para ítems en común; y el sector restante corresponderá a los ítems únicos. En esta etapa del diseño se identificó un problema de incompatibilidad entre la estrategia targeting designada y la calibración concurrente.

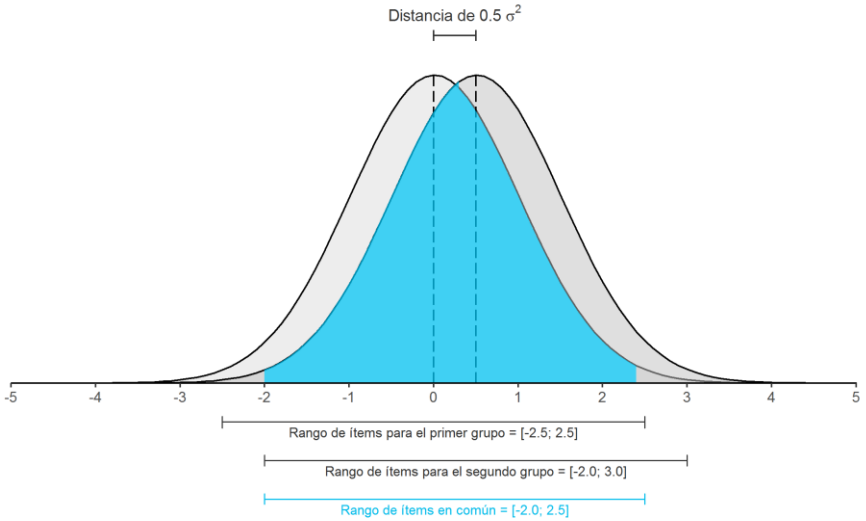

Figura 4.4. Intersección de las distribuciones de habilidad cuando la distancia entre medias es de 0.5 desviaciones estándar.

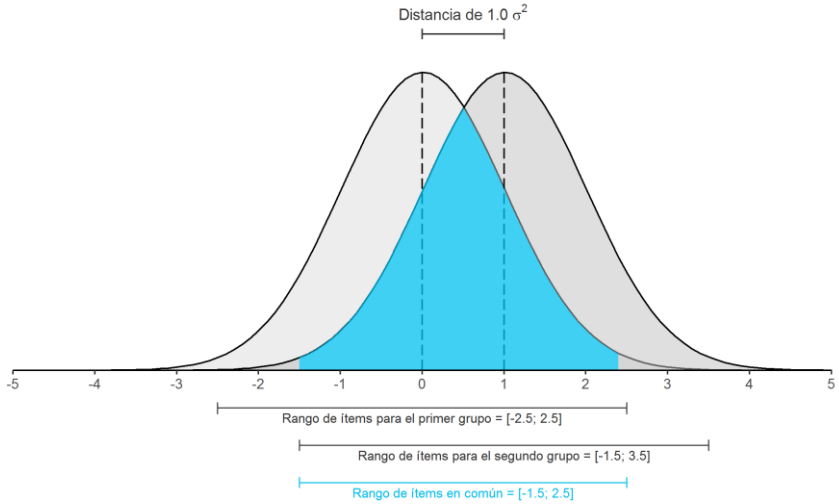

Figura 4.5. Intersección de las distribuciones de habilidad cuando la distancia entre medias es de 1.0 desviaciones estándar.

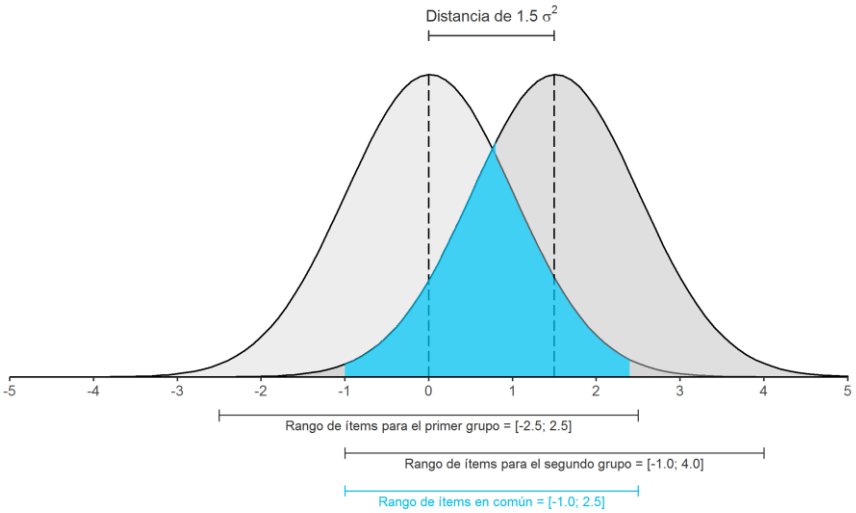

Figura 4.6. Intersección de las distribuciones de habilidad cuando la distancia entre medias es de 1.5 desviaciones estándar.

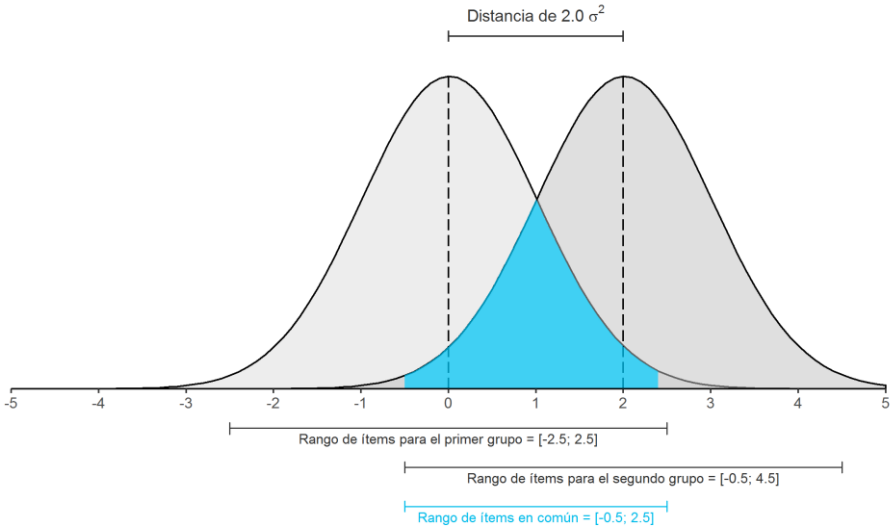

Figura 4.7. Intersección de las distribuciones de habilidad cuando la distancia entre medias es de 2.0 desviaciones estándar.

Como se puede observar en las figuras 4.4 a 4.7, en cada condición del experimento se puede designar un rango de ítems en común en función a la intersección entre las distribuciones de habilidad de ambos grupos. No obstante, los gráficos también permiten observar que conforme la distancia

entre las medias aritméticas aumenta, menor es el rango disponible para designar ítems en común para ambas pruebas. En otras palabras, si se desea designar ítems en común y, al mismo tiempo, mantener una estrategia de targeting de -2.5 a 2.5 desviaciones estándar alrededor de la media aritmética de las distribuciones de habilidad para cada grupo, el rango disponible para ítems en común se encontrará condicionado a la intersección entre las distribuciones de habilidad de los grupos.

El problema asociado a un rango limitado de dificultad para los ítems en común es que un proceso de equiparación idóneo requiere que dichos ítems representen apropiadamente a todos los niveles del rasgo latente (Kolen y Brenan, 2014). DeMars (2002) estudió el efecto de este diseño en el contexto de un escalamiento vertical para dos grupos, en donde los ítems en común representaban los ítems más difíciles para el primer grupo y los más fáciles para el segundo grupo, una situación equivalente a la presentada en este estudio.

Entre sus hallazgos, el autor encontró que el diseño resultaba en la sobreestimación de los parámetros de dificultad de los ítems de la forma con menor dificultad; y la subestimación de los parámetros de los ítems de la forma con mayor dificultad. Ante esto, Gonzales y Wiberg (2017) analizaron los resultados del estudio de simulación de DeMars (2002), y se dieron cuenta de que el sesgo de la estimación fue pequeño y; por lo tanto, no implica diferencias relevantes en la práctica. En consecuencia, se optó por mantener el diseño de bloques, tomando en cuenta que puede implicar un sesgo pequeño en la estimación de parámetros.

Como la situación en donde la diferencia de medias aritméticas es de 2.0 desviaciones estándar presenta la menor intersección entre las distribuciones de habilidad los grupos (figura 4.7), el rango de dicha intersección se delimitó como el rango de dificultad para los ítems en común de todas las condiciones del experimento. Esto significa que todas las condiciones tendrán los mismos ítems en común en el rango de dificultad de -0.5 a 2.5.

Una vez justificado el uso del rango de dificultad para los ítems en común, el siguiente paso para asegurar un proceso de equiparación idóneo fue designar la cantidad de dichos ítems. Kolen y Brenan (2014) afirman que no existe un consenso en relación a la cantidad adecuada de ítems en común que asegure la efectividad de la equiparación; sin embargo, los expertos sugieren emplear una cantidad entre 20% a 30% del total de ítems de la prueba. Por estos motivos, se optó por designar 16 ítems en común, que representan un 40% del total de ítems de las dos pruebas on target para los grupos de la población.

Al designar el número de ítems en común y su respectivo rango de dificultad, solo resta determinar la distribución de los parámetros de dificultad de los ítems. Burga (en prensa), encontró que la interacción entre la distribución de la habilidad de las personas y la distribución de la dificultad de los ítems no influye en la confiabilidad de las puntuaciones. En otras palabras, tener una distribución de la dificultad de los ítems distinta a la de la distribución de las personas no implica una disminución en la precisión de las puntuaciones.

Estos resultados suponen una mayor flexibilidad sobre los supuestos distribucionales de la dificultad de los ítems. A pesar de ello, Wright y Stone (1999) sugieren que los ítems deben tener una separación apropiada a lo largo del continuum latente para obtener una mayor precisión de la estimación de parámetros. Por esta razón, los parámetros de dificultad de los ítems fueron generados a través de la función *seq()* del programa base de R (versión 3.5.2; R Core Team, 2018), la cual permite generar secuencias equidistantes entre dos números preestablecidos. De esta manera, se maximiza la separación entre parámetros de dificultad de los 16 ítems en común.

Los 24 ítems restantes de cada prueba fueron designados como ítems únicos. Debido a que la media aritmética del primer grupo se mantiene constante en todas las instancias del experimento (ver tabla 4.3), el rango disponible para los parámetros de dificultad de los ítems únicos también es el mismo para cada condición simulada, entre -2.5 y -0.5. En contraste, la

media aritmética del segundo grupo sí varía de acuerdo a los niveles de la variable independiente distancia entre medias; por ello, el rango disponible para establecer parámetros de dificultad de los ítems únicos también varió, los rangos de ambos grupos se presentan en la tabla 4.6.

Tabla 4.6

Ítems únicos para los grupos que componen la población

Distancia entre medias aritméticas	Conjunto de ítems únicos para el primer grupo	Conjunto de ítems únicos para el segundo grupo
0.5	$[-2.5;\ -0.5]$	$[-2.0;\ -0.5] \cup [2.5; 3.0]$
1.0	$[-2.5;\ -0.5]$	$[-1.5;\ -0.5] \cup [2.5; 3.5]$
1.5	$[-2.5;\ -0.5]$	$[-1.0;\ -0.5] \cup [2.5; 4.0]$
2.0	$[-2.5;\ -0.5]$	$[2.5; 4.5]$

La designación de los parámetros de dificultad de los ítems únicos también se realizó a través de la función *seq()* del programa base de R (versión 3.5.2; R Core Team, 2018) para mantener la congruencia con la recomendación de Wright y Stone (1999) mencionada anteriormente. Para mayor información sobre los parámetros de ítems generados en cada condición, el código de R utilizado en el experimento puede ser consultado en el apéndice 1. Además, como síntesis de este punto, los rangos de dificultad de los ítems únicos y en común para cada condición del experimento se presentan en la figura 4.8.

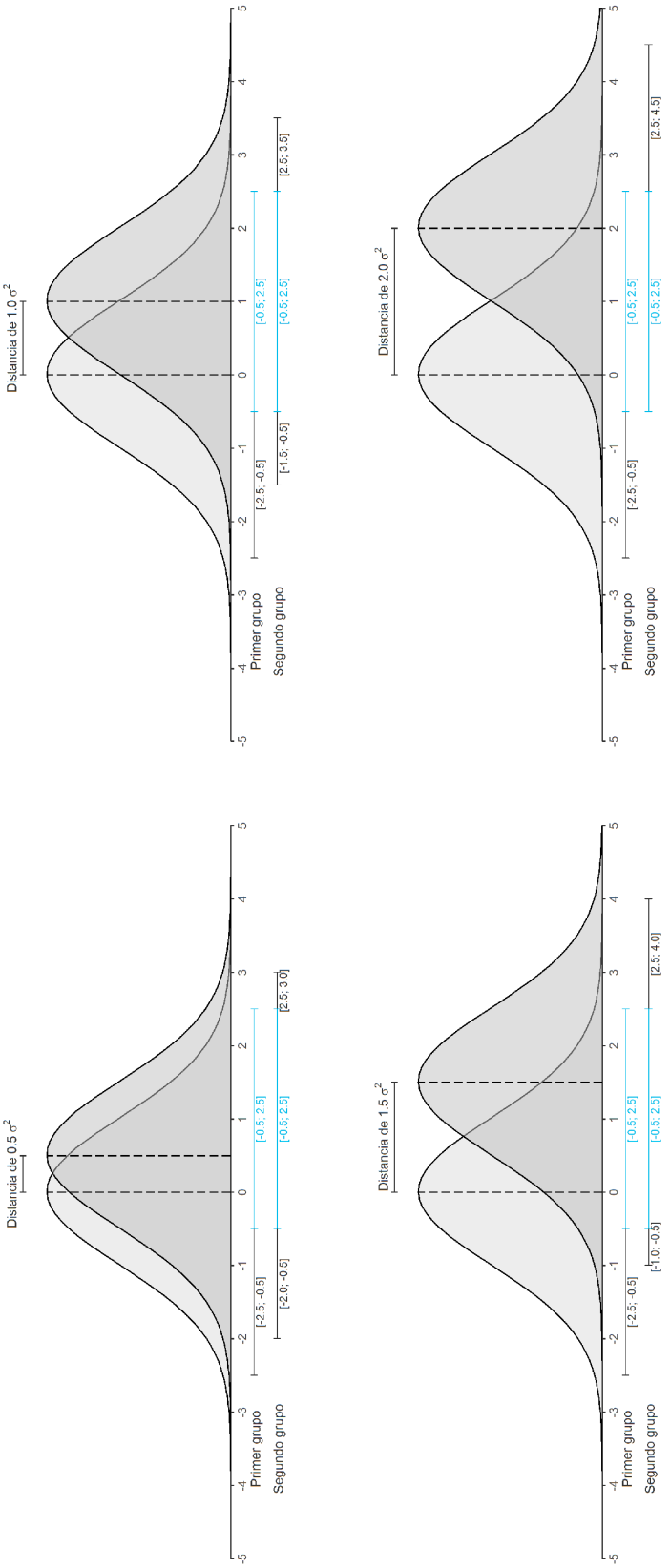

Figura 4.8. Síntesis de los rangos de dificultad de los ítems únicos y en común para los dos grupos de la población.

4.4 Procedimiento de recolección de datos

La complejidad del diseño experimental propuesto en esta investigación requiere un grado sumamente alto de control sobre distintas variables. Desafortunadamente, los métodos tradicionales de recolección de datos (i.e., la evaluación a través de instrumentos psicométricos) no permitirían ejercer dicho grado de control sobre condiciones tan complejas como establecer una distribución de parámetros de personas e ítems congruente con una estrategia targeting designada.

Afortunadamente, existe una aproximación matemática que permite superar estas limitaciones, el método de simulaciones Monte Carlo. Históricamente, diversas disciplinas científicas han utilizado simulaciones en la investigación y resolución de problemas, incluyendo a la psicometría (Harwell, 1997; Olvera, 2017). Como se mencionó anteriormente, en esta sección se delimitarán las fases del experimento de simulaciones correspondientes a la generación de datos aleatorios, la estimación de los parámetros del modelo Rasch, y la estimación de la confiabilidad en cada condición del experimento.

Sobre la base de los supuestos distribucionales establecidos en las secciones anteriores, los datos aleatorios generados en este estudio son respuestas a ítems dicotómicos congruentes con el modelo Rasch. Para generar una respuesta, se utilizó la función *obs.resp()* creada por el área de psicometría de la UMC, cuya metodología consiste en integrar un vector de parámetros de habilidad y un vector de parámetros de dificultad para estimar la probabilidad de acierto en cada interacción entre personas e ítems sobre la base del modelo Rasch para ítems dicotómicos. Después, esta probabilidad se establece como el parámetro p de un ensayo Bernoulli, un experimento aleatorio con dos posibles resultados: acierto o fallo (Grami, 2020).

El resultado del experimento Bernoulli (o experimento Binomial con $n = 1$) se realiza en cada interacción entre personas e ítems, de modo que el *output* final es una matriz de datos con ceros y unos que representan respuestas a ítems dicotómicos (para mayor detalle sobre esta función, consultar el apéndice 1). Es importante mencionar que la función de la

UMC produce el mismo resultado que las funciones *rmvlogis()* del paquete ltm (Rizopoulos, 2018, versión 1.1-1) y *sim.irt()* del paquete *psych* (Revelle, 2019, versión 1.8.12). Ambos paquetes son altamente reconocidos por su utilidad para el análisis psicométrico.

Posteriormente, la base de datos fue ajustada al modelo Rasch para ítems dicotómicos con el objetivo de estimar los parámetros de habilidad y dificultad. La estimación de los parámetros de dificultad de los ítems se realizó a través del método Marginal Maximum Likelihood Estimation [MMLE] (Bock y Lieberman, 1970). A diferencia de otras técnicas iterativas, el MMLE permite la estimación de parámetros incluso cuando existe información insuficiente para estimar a algunos individuos; por ejemplo, puntajes extremos en donde se acertó o falló a todos los ítems; además, es robusto ante valores perdidos (Linacre, 1999).

En estudios de simulación se ha demostrado que el MMLE es superior a otros métodos tradicionales como el JMLE en diversas condiciones (Drasgow, 1989; Lim y Drasgow, 1990; Yen, 1987). Esto ocurre porque el algoritmo estima los parámetros de los ítems excluyendo a los parámetros de personas de la ecuación de verosimilitud (Tinsley y Brown, 2000). En su lugar, se asume un supuesto distribucional sobre la habilidad de las personas, usualmente una distribución normal (Linacre, 1999). Este procedimiento se realizó a través de la función *tam.mml()* del paquete Test Analysis Modules [TAM] (versión 3.3-10; Robitzsch, Kiefer y Wu, 2019).

Como el MMLE estima únicamente los parámetros de dificultad de los ítems, se requiere de un método complementario para estimar los parámetros de habilidad de las personas. En este estudio se empleó el método Expected A Posteriori [EAP] (Bock y Aitkin, 1981), una aproximación bayesiana que contrasta una distribución *anterior* (usualmente una normal o uniforme) con los datos observados con el objetivo de estimar una distribución *posterior* para cada patrón de respuestas (Mislevy y Stocking, 1989). Posteriormente, el algoritmo del EAP utiliza la media aritmética de la distribución posterior de cada patrón

de respuestas para asignar un parámetro de habilidad a cada persona (Engelhard y Wind, 2018).

El EAP presenta una serie de ventajas importantes con respecto a otros estimadores (Chen, Hou y Dodd, 1998; Wang y Wang, 2001). En términos computacionales, el EAP es más eficiente que otros métodos utilizados en la estimación de parámetros de habilidad como el WLE, MLE o MAP; debido a que sus cálculos implican un menor número de operaciones (Bock y Mislevy, 1982). Además, el EAP permite estimar parámetros de habilidad para personas con patrones de respuesta en donde se acierta o falla a la totalidad de ítems, una limitación importante de los métodos WLE y MLE (Bock y Mislevy, 1982).

Aunque algunos autores afirman que la precisión de las estimaciones del EAP depende de la idoneidad de la distribución anterior establecida (Mislevy y Stocking, 1989; Seong, 1990), diversos estudios de simulaciones han demostrado que las estimaciones del EAP son tan precisas y estables como las del MLE, WLE y MAP (Wang y Wang, 2001; Bock y Mislevy, 1982), incluso si la distribución anterior no es congruente con la distribución de la variable latente (Chen et al., 1998; Chen, Hou, Fitzpatrick y Dodd, 1997).

Afortunadamente, al estimar los parámetros de dificultad de los ítems con la función *tam.mml(),* la última iteración automáticamente realiza el cálculo de parámetros de habilidad utilizando el estimador EAP. Al obtener dichos parámetros, ya es posible estimar el coeficiente de confiabilidad de separación de personas, el cual representa el indicador de la variable dependiente del estudio. Este coeficiente requiere como insumo únicamente la varianza de los parámetros de habilidad y la media cuadrática de sus errores estándar (Bond y Fox, 2015). Aunque su cálculo es relativamente sencillo, *tam.mml()* también estima este indicador; por lo tanto, no fue necesario emplear funciones complementarias.

De acuerdo a la metodología Monte Carlo, el procedimiento descrito, desde la generación de matrices de respuestas a ítems dicotómicos hasta la estimación del índice de confiabilidad de separación de personas debe ser

replicado un número determinado de veces para cada condición del estudio. En la literatura no existe un consenso delimitado en relación al número de réplicas apropiado para un estudio de simulaciones. Por ejemplo, Harwell et al. (1996) sugiere un mínimo de 25 réplicas para experimentos que involucren modelos TRI; mientras que, Mundform, Schaffer, Kim, Shaw y Thongteeraparp (2011) concluyen que 5000 réplicas deberían ser suficientes para producir resultados estables. Ambas posturas ejemplifican la diversidad de recomendaciones presentes en la literatura especializada.

Es importante reconocer que el número de réplicas también debe considerar la cantidad de factores utilizados en el estudio y la capacidad computacional con la que se dispone (Feinberg y Rubright, 2016). Este experimento incorpora ocho factores de estudio, y el dispositivo en donde se realizó el procedimiento es una laptop con un procesador Intel ® Core ™ i5-7300HQ CPU @ 2.50GHz, con una memoria RAM de 12.0 GB y un sistema operativo Windows 10 de 64 bits. Considerando ambos aspectos, se decidió implementar 1000 réplicas para cada condición del experimento, lo que resulta en un total de 8000 iteraciones del proceso previamente descrito.

Continuando con los pasos recomendados para implementar un experimento de simulaciones Monte Carlo, Feinberg y Rubright (2016) consideran que, al finalizar el proceso iterativo del experimento, los parámetros estimados en todas las réplicas deben ser evaluados sobre la base de distintos criterios que delimitan la precisión de la recuperación de parámetros verdaderos, dichos criterios se presentan en la tabla 4.7.

Tabla 4.7

Métodos para contrastar los parámetros estimados y los parámetros verdaderos

Método	Descripción	Fórmula
Bias	Provee una medida del promedio de la distancia entre el parámetro estimado y el parámetro verdadero.	$\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_{Verdadero})}{n}$
SE	Es la desviación estándar del parámetro estimado entre las réplicas, una medida del error aleatorio.	$\sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \bar{\theta})^2}{n-1}}$
MSE	Mide el promedio de los cuadrados de las desviaciones entre los parámetros estimados y los parámetros verdaderos.	$\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_{Verdadero})^2}{n-1}$
RMSE	Es la raíz cuadrada del MSE	$\sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_{Verdadero})^2}{n-1}}$
MAD	Provee una magnitud absoluta del promedio de la distancia entre el parámetro estimado y el parámetro verdadero.	$\frac{\sum_{i=1}^{n}\lvert\widehat{\theta_i} - \theta_{Verdadero}\rvert}{n-1}$

Nota. Adaptado de "Conducting simulation studies in psychometrics," por R. A. Feinberg y J. D. Rubright, 2016, *Educational Measurement Issues and Practice, 35*(2), p. 40. Copyright 2016 por The National Council on Measurement in Education.

Usualmente, estos criterios son utilizados cuando el objetivo de las simulaciones es la recuperación de parámetros verdaderos (Bulut y Sünbül, 2017). Sin embargo, como este no es el objetivo de este experimento, no se realizaron dichas estimaciones. En su lugar, el presente experimento se enfoca en el análisis de los coeficientes de confiabilidad de separación de personas estimados y en su comparación en función a distintas situaciones de diferencias entre medias aritméticas de la habilidad de dos grupos que componen una población.

Por estos motivos, la programación del experimento se realizó de modo que la confiabilidad estimada en cada una de las 8000 iteraciones sea almacenada en una base de datos organizada en función a la condición del experimento de donde se obtuvo dicha observación. Esta base de datos representa el resultado final del proceso de recolección de datos del experimento. El código de programación desarrollado para estas fases se adjunta en el apéndice 1 con el objetivo de fomentar la reproducibilidad y replicabilidad de los resultados de este estudio, tal y como se recomienda en la literatura especializada (Stodden, Guo y Ma, 2013).

CAPÍTULO V: RESULTADOS

De acuerdo a Harwell et al. (1996), la fase final de un experimento Monte Carlo consiste en analizar los resultados obtenidos en las simulaciones a través de métodos descriptivos e inferenciales congruentes con los objetivos de la investigación. En este capítulo se presentan los resultados de los respectivos análisis junto con argumentos que justifican la elección de cada método implementado.

5.1 Datos sobre el proceso de simulación

La simulación de datos fue implementada en una laptop con un procesador Intel ® Core ™ i5-7300HQ CPU @ 2.50GHz. El dispositivo contó con una memoria RAM de 12.0 GB y un sistema operativo Windows 10 de 64 bits. A través de la función *sys.time()* del programa base de R (versión 3.5.2; R Core Team, 2018), se registró un tiempo aproximado de 2 horas y 45 minutos. El código de dicho procedimiento puede ser consultado en el apéndice 1.

5.2 Análisis exploratorio de datos

De acuerdo a John W. Tukey (1997), antes de implementar cualquier análisis estadístico inferencial, es indispensable realizar un análisis exploratorio de datos. En esta fase preliminar, Pearson (2018) recomienda estimar una serie de estadísticos descriptivos que brinden información acerca de la tendencia y dispersión de los datos. Además, dichos indicadores deben ser congruentes con la naturaleza de la variable de estudio (Howitt y Cramer, 2017). Los valores estimados del coeficiente de confiabilidad de separación de personas pueden considerarse como variables continuas en una escala de medición de intervalo. En este tipo de variables, las diferencias de la misma magnitud numérica tienen la misma distancia (Heumann y Schomaker, 2016; Ho, 2018). Por ejemplo, la

distancia entre una confiabilidad de .80 a .81 es la misma que una confiabilidad de .90 a .91. Por tal motivo, es posible utilizar estadísticos que suponen propiedades de equidistancia en los datos como la media aritmética y la desviación estándar (Stevens, 1946). Estos estadísticos son presentados en la tabla 5.1 junto a otros indicadores complementarios como los intervalos de confianza de las medias aritméticas, las medias aritméticas recortadas, y los valores mínimos y máximos.

.

Tabla 5.1

Estadísticos descriptivos de la confiabilidad estimada en el experimento

Target	Distancia entre las medias	*n*	*M*	95% IC	*DE*	Media recortada	Valor Mínimo	Valor Máximo
Grupos	0.5	1000	.860	[.856, .866]	.003	.860	.852	.868
	1.0	1000	.877	[.872, .881]	.002	.877	.868	.884
	1.5	1000	.897	[.893, .901]	.002	.897	.889	.903
	2.0	1000	.915	[.912, .918]	.002	.915	.910	.920
Población	0.5	1000	.867	[.862, .872]	.002	.867	.859	.876
	1.0	1000	.883	[.879, .887]	.002	.883	.876	.890
	1.5	1000	.902	[.899, .905]	.002	.902	.897	.907
	2.0	1000	.919	[.917, .921]	.001	.919	.915	.922

Nota. La media recortada representa la media aritmética obtenida al eliminar el 10% de datos con valores más altos y el 10% con valores más bajos. n = número de réplicas, M = media aritmética, DE = desviación estándar, IC = intervalos de confianza.

Feinberg y Rubright (2016) recomiendan complementar estos indicadores con estrategias gráficas que faciliten la presentación de los resultados de un experimento Monte Carlo. Particularmente, se optó por generar un *boxplot* (también denominado *gráfico de cajas y bigotes* en español) de los coeficientes de confiabilidad de separación de personas obtenidos en cada condición del experimento. Este gráfico es especialmente útil para detectar la presencia de valores extremos y analizar la dispersión, y forma de los datos (Field, 2018; Pearson, 2018).

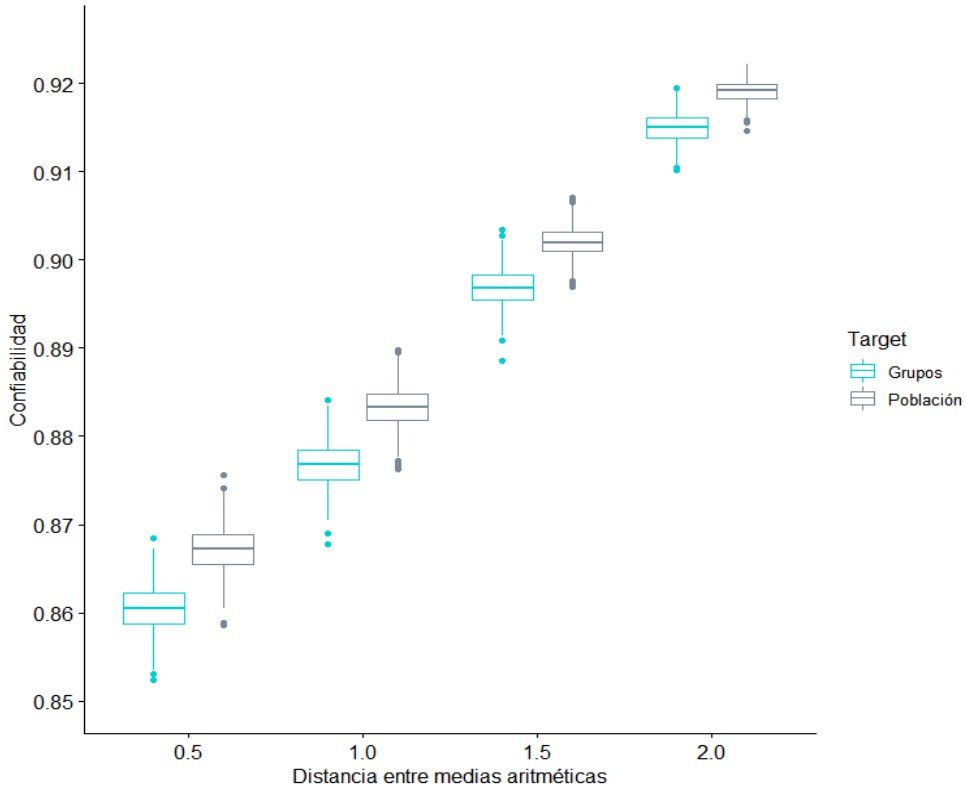

Figura 5.1. Gráfico de cajas y bigotes de la confiabilidad estimada en cada condición del experimento.

En la figura 5.1 es posible apreciar que, conforme la distancia entre las medias aritméticas de habilidad de ambos grupos aumenta, la confiabilidad también incrementa al considerar como target tanto a los grupos como a la población. Sin embargo, en los cuatro escenarios de distancia entre medias aritméticas de habilidad de los grupos, se observa que la confiabilidad estimada en las evaluaciones con un target en la población es mayor a la estimada al considerar como target a los grupos que la componen.

Finalmente, es posible identificar la presencia de datos atípicos en todos los escenarios; estos valores *outliers* pueden sesgar los resultados de los modelos estadísticos (Wiley y Wiley, 2019). En la literatura metodológica existen diversas aproximaciones para lidiar con los valores extremos como *el adjusted predicted value*, *studentized deleted residual* [SDR] o *Cook's distance*, cada uno con valores recomendados como puntos de corte que pueden ser utilizados para eliminar casos que puedan distorsionar los

análisis. No obstante, en el presente estudio se utilizarán métodos de estimación robustos que utilizan la media recortada como estadístico base. La media recortada es la media aritmética de los valores estimada al descartar un determinado porcentaje (usualmente 10%) de casos en los extremos de la distribución.

5.3 Análisis inferencial de datos

El análisis de varianza [ANOVA], desarrollado por Ronald A. Fischer en 1930, fue la estrategia de análisis estadístico seleccionada para la comparación de la confiabilidad estimada en los grupos formados por la interacción entre las variables independientes target y distancia entre medias aritméticas. En general, el ANOVA representa a una familia de procedimientos estadísticos que utilizan la prueba *F* para evaluar el ajuste de un modelo lineal a los datos observados (Field, 2007). También, se asocia con el análisis de diseños de investigación experimentales en donde una variable independiente categórica es manipulada para evaluar su efecto en una variable dependiente continua (Wahed y Tang, 2010).

El resultado final de la prueba *F* es una evaluación general de las diferencias de medias de grupos entre los niveles de la variable independiente categórica (Field, 2007). Cuando el análisis involucra más de una variable independiente, se denomina *ANOVA factorial*. En este caso, múltiples pruebas *F* son empleadas para contrastar los *efectos principales* de las variables independientes sobre la variable dependiente y para evaluar el *efecto de interacción* para cada combinación de efectos principales (Field, 2007; Leppink, 2019).

5.3.1 Evaluación de supuestos del ANOVA factorial

Al igual que la mayoría de pruebas estadísticas, esta técnica se desarrolla sobre la base de una serie de supuestos preestablecidos (Leppink, 2019, Paolella, 2019). El incumplimiento de dichos supuestos implica resultados sesgados; por lo tanto, es necesario evaluar cada uno de ellos (Denis, 2019).

En principio, variables independientes tienen que ser categóricas y la variable dependiente debe encontrarse en un nivel de medición de intervalo (Wahed y Tang, 2010). Dado el diseño de este experimento previamente descrito, ambas condiciones preliminares se cumplen.

Antes de introducir el resto de supuestos del modelo, es importante mencionar que las técnicas de análisis de varianza forman parte de una familia de análisis estadísticos denominados Modelos Lineales Generalizados [GLM, por sus siglas en inglés] (Field, 2018). Rutherford (2011) relata que, históricamente, la terminología empleada para referirse a los supuestos del ANOVA era distinta a la utilizada frente a los supuestos del GLM; mientras que una hacía referencia a las puntuaciones en cada condición del experimento, la otra se enfocaba en los errores del modelo, respectivamente. A pesar de ello, ambas notaciones son equivalentes.

Tabla 5.2

Los supuestos del ANOVA y GLM

N°	Supuestos de ANOVA	Supuestos de GLM
1	Cada condición contiene una muestra aleatoria de la población.	Cada condición contiene una muestra aleatoria de la población.
2	Las puntuaciones en cada condición son independientes.	Los errores son independientes.
3	Las puntuaciones en cada condición se distribuyen normalmente.	Los errores se distribuyen normalmente.
4	Las varianzas de las puntuaciones en cada condición del experimento son homogéneas.	Los errores son homoscedásticos (los errores exhiben varianza común a través de todos los valores de las variables predictoras)

Nota. Adaptado de *ANOVA and ANCOVA. A GLM approach* (2nd ed., p. 235), por A. Rutherford, 2011, Hoboken, NJ: John Wiley & Sons. Copyright 2011 por John Wiley & Sons, Inc.

En la tabla 5.2 se presenta la equivalencia entre las terminologías tradicionales de los supuestos del ANOVA y la taxonomía de los GLM. El primer supuesto de la técnica se cumple debido a la naturaleza propia de los estudios de simulaciones. En efecto, el método Monte Carlo implica un muestreo aleatorio de datos sobre la base de modelos distribuciones (Feinberg y Rubright, 2016); por lo tanto, la aleatoriedad de las muestras queda garantizada.

Por su parte, el supuesto de independencia de los errores del modelo es uno de los más importantes y más difíciles de contrastar empíricamente. De acuerdo a Denis (2019), el propio diseño del estudio y la manera en que los datos fueron recolectados aseguran el cumplimiento de este supuesto. Afortunadamente, en esta investigación, la programación del experimento se ha desarrollado de modo que la simulación de un caso particular no presente algún tipo de dependencia frente al resto.

Los últimos dos supuestos requieren de un contraste empírico con los datos obtenidos en las simulaciones. Con respecto al supuesto de normalidad, al realizar un análisis empleando el estadístico *F*, lo recomendable es comprobar que las observaciones de la variable dependiente obtenidas en cada uno de los grupos, se distribuyan normalmente a nivel poblacional (Field, 2018). Esto es equivalente a que los errores del modelo se distribuyan normalmente, en la terminología GLM (Rutherford, 2011). Existen distintas maneras para evaluar este supuesto, en esta investigación se empleó la prueba de Shapiro-Wilk, la cual ha demostrado tener mayor potencia que el método tradicional de Kolmogorov-Smirnov, incluso al aplicar la corrección de Lilliefors (Ghasemi y Zahedias, 2012). La prueba de normalidad se realizó a través de la función *shapiro.test()* del paquete *stats* preinstalado en el programa base de R (versión 3.5.2; R Core Team, 2018). Los resultados de este procedimiento se presentan en la tabla 5.3.

Tabla 5.3

Pruebas de normalidad para cada condición del experimento

Target	Distancia entre las medias	*W*	*df*	Significancia estadística
Grupo	0.5	.999	1000	.707
	1.0	.998	1000	.143
	1.5	.999	1000	.663
	2.0	.997	1000	.118
Población	0.5	.999	1000	.982
	1.0	.998	1000	.475
	1.5	.998	1000	.187
	2.0	.998	1000	.259

Nota. *W* hace referencia al valor obtenido del estadístico Shapiro-Wilk.

En general, los resultados indican que no existen evidencias empíricas suficientes para descartar que la confiabilidad estimada en cada uno de los grupos se distribuya normalmente a nivel de la población. Un resultado similar puede observarse al analizar los datos a través de un Q-Q plot, en donde se comparan los cuantiles observados en los datos de cada grupo con los cuantiles de una distribución teórica esperada, en este caso, la distribución normal (Grami, 2020; Mishra et al., 2019; Wiley y Wiley, 2019). En la figura 5.2 se observa que, en todos los casos, se presenta una superposición de los datos empíricos, representados por los puntos, con la distribución teórica normal, representada por la línea diagonal.

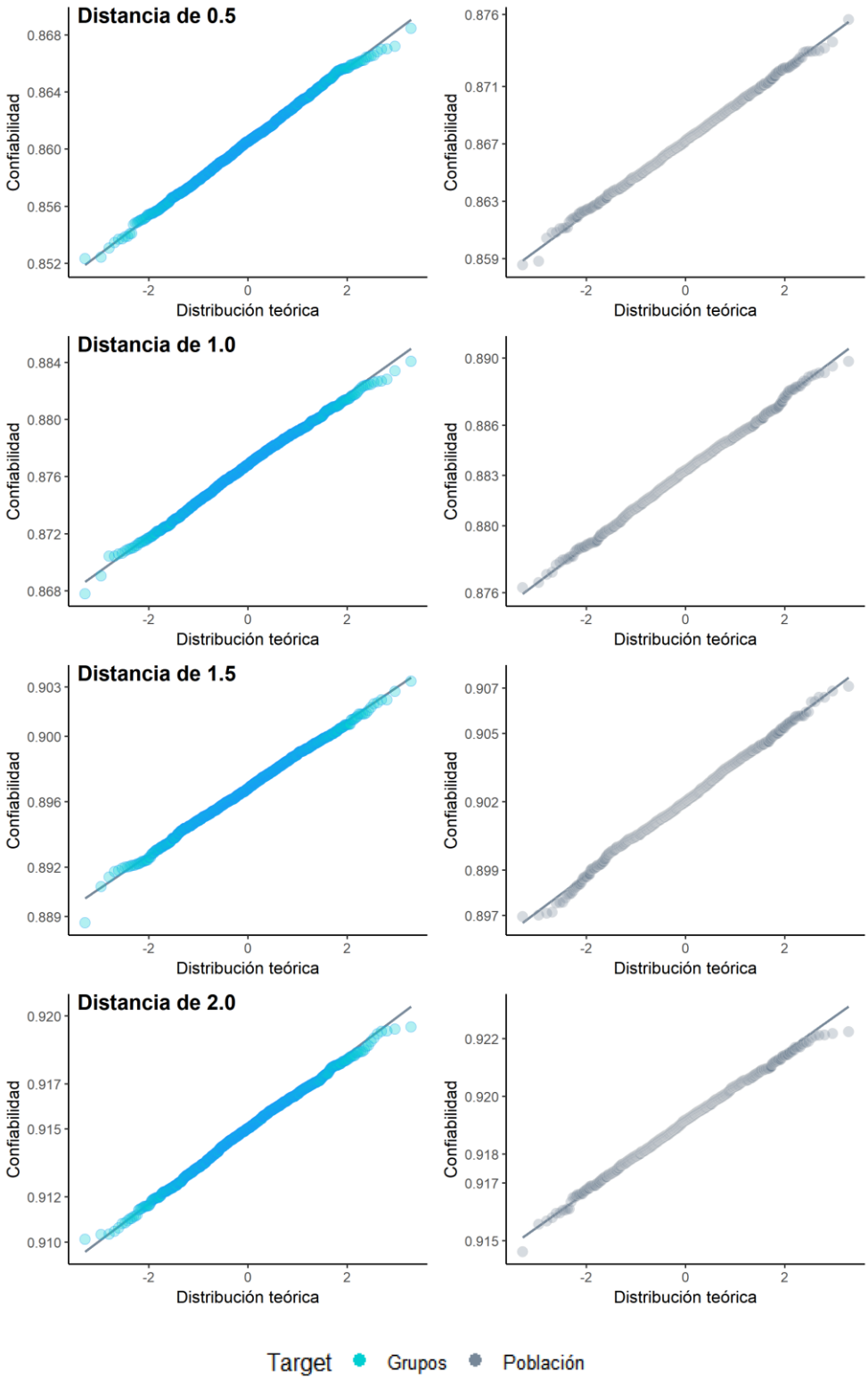

Figura 5.2. Gráficos Q-Q para cada condición del experimento.

El último supuesto que subyace a esta técnica es la homogeneidad de varianzas, cuyo cumplimiento implica que las varianzas en todos los grupos a ser comparados deben ser iguales (Field, 2018). Esto es equivalente a que exista *homoscedasticidad* en los errores del modelo, de acuerdo a la terminología GLM (Rutherford, 2011). Al igual que el supuesto de normalidad, la homogeneidad de varianzas puede ser contrastada a través de diversos métodos. En este experimento, se utilizó la prueba de Levene a través de la función *leveneTest()* del paquete *car* (versión 3.0-5; Fox y Weisberg, 2019) en R (versión 3.5.2; R Core Team, 2018).

Tabla 5.4

Resultados de la prueba de homogeneidad de varianzas de Levene

Centrado	*F*	*df1*	*df2*	Significancia estadística
Media	106.67	7	7992	<.001*
Mediana	106.50	7	7992	<.001*

Nota. El diseño considera la intersección entre las variables target y distancia.

*p<.05.

Los resultados de este procedimiento (presentados en la tabla 5.4) indican que existen diferencias estadísticamente significativas entre las varianzas de la confiabilidad estimada entre los ocho grupos de estudio al considerar un centrado en la media $F(7{,}7992) = 106.67, p < .001$; y un centrado en la mediana $F(7{,}7992) = 106.50, p < .001$.

Es importante considerar que, cuando se tiene un tamaño de muestra grande, la potencia del test de Levene aumenta. En otras palabras, el test será capaz de detectar diferencias muy pequeñas en las varianzas de los grupos de estudio; por ello, lo más probable es que el resultado sea estadísticamente significativo (Greenland et al., 2016).

En estas circunstancias es recomendable emplear otro tipo de pruebas que complementen los resultados obtenidos en el test de Levene (Field, Miles y Field, 2012). Por estos motivos, se optó por realizar el estadístico F_{max} de

Hartley, también conocido como el ratio de varianzas. Tal y como su nombre indica, este procedimiento consiste en un ratio entre la varianza más grande y la más pequeña entre los grupos. Para esto, se empleó la función *hartleyTest()* del paquete *PMCMRplus* (versión 1.4.2; Pohlert, 2019) en R (versión 3.5.2; R Core Team, 2018).

Tabla 5.5

Resultados del test de Hartley

F_{max}	*df*	Significancia estadística
4.910	999	<.001*

Nota. F_{max} representa al valor obtenido en el ratio de varianzas de Hartley.

*p<.05.

La tabla 5.5 presenta los resultados de este procedimiento, los cuales indican que existen diferencias estadísticamente significativas entre las varianzas de la confiabilidad estimada entre los grupos de estudio $F_{max}(999) = 4.9102, p < .001$. Estos resultados son congruentes con aquellos obtenidos en la prueba de Levene; por estos motivos, se concluye que no se cumple el supuesto de homogeneidad de varianzas.

5.3.2 ANOVA factorial robusto

Como se mencionó anteriormente, el incumplimiento de los supuestos de cualquier método estadístico supone la imprecisión de sus estimaciones (Hoekstra, Kiers y Johnson, 2012). En este caso, el incumplimiento del supuesto de homogeneidad de varianzas puede ocasionar que las estimaciones del ANOVA factorial se encuentren sesgadas (Denis, 2019). Por un lado, existe la posibilidad de que el estadístico sea conservador, es decir, es más probable que produzca un resultado que no sea estadísticamente significativo cuando sí existen diferencias genuinas en la población. Por otro lado, la técnica también puede ser liberal; en otras

palabras, presentará resultados estadísticamente significativos cuando en realidad no hay diferencias entre los grupos de la población (Field, 2018).

Ante esto, Field (2018) afirma que el ANOVA es apropiadamente robusto ante la violación al supuesto de homogeneidad de varianzas cuando los tamaños de muestra en cada condición del experimento son iguales. Sin embargo, en la literatura especializada existen diversas alternativas ante el ANOVA factorial tradicional, específicamente para las situaciones en donde no se cumple la homoscedasticidad en los datos. Particularmente, Rand Wilcox (2012; 2017a; 2017b) ha desarrollado una amplia gama de versiones robustas de técnicas de análisis estadístico tradicionales, entre ellos, el análisis de varianza factorial robusto.

La versión Wilcox (2012) del ANOVA factorial es robusta ante la violación del supuesto de homogeneidad de varianzas y, a diferencia del ANOVA tradicional, esta técnica compara las medias aritméticas recortadas. Este procedimiento se desarrolló a través de la función *t2way()* del paquete *WRS2* (versión 1.0-0, Mair y Wilcox, 2019a) en R (versión 3.5.2; R Core Team, 2018). Para una guía detallada sobre cómo aplicar este análisis y otros estadísticos robustos consultar Mair y Wilcox (2019b).

Tabla 5.6

Resultados del ANOVA factorial robusto basado en medias recortadas

Variable	*F*	Significancia estadística
Target	13842.322	<.001*
Distancia	786955.515	<.001*
Interacción entre target y distancia	571.561	<.001*

Nota. La media aritmética recortada se estimó excluyendo al 10% de observaciones ubicados en el extremo superior de la distribución y al 10% de las observaciones del extremo inferior. Además, *t2way()* no reporta grados de libertad porque utiliza un valor crítico ajustado.

*p<.05.

Los resultados del análisis de varianza factorial (presentados en la tabla 5.6), a través del procedimiento basado en medias recortadas de Wilcox (2017a; 2017b), señalan que existe un efecto estadísticamente significativo del target empleado, $F = 13842.322\,, p < .001$, la distancia entre las medias aritméticas de los grupos poblacionales $F = 786955.515\,, p < .001$ y de la interacción entre ambas variables $F = 571.561, p < .001$.

5.3.3 Comparaciones múltiples

Al interpretar estos resultados se debe tomar en cuenta que el ANOVA factorial es una prueba ómnibus que indica si alguno de los factores o su interacción contribuyen a la predicción de la variable dependiente (Rutherford, 2011). Al haber encontrado efectos estadísticamente significativos por parte de las variables independientes y su interacción, el siguiente paso para alcanzar el objetivo del estudio es comparar directamente la confiabilidad obtenida a través de tests on target para dos grupos con aquella obtenida a partir de un solo test para toda la población.

Dicha comparación puede ser fácilmente realizada a partir de pruebas de comparaciones múltiples, un conjunto de análisis que complementan al ANOVA (Denis, 2019). Entre todas las alternativas disponibles en la literatura, se optó por utilizar el estadístico Ψ (Wilcox, 2012). Wilcox (2017b) desarrolló esta prueba específicamente para complementar el ANOVA factorial robusto basado en medias recortadas (Mair y Wilcox, 2019b).

El estadístico Ψ propuesto por Wilcox (2012) se basa en la extensión del método de Yuen (1974) para la comparación de grupos a partir de las medias recortadas, en conjunto con una generalización del procedimiento T3 de Dunnett (1980) empleado como estrategia para las comparaciones múltiples cuando existe heterocedasticidad en los datos. Dicho procedimiento se realizó a través de la función *lincon()* del paquete *WRS2* (versión 1.0-0, Mair y Wilcox, 2019a) en R (versión 3.5.2; R Core Team, 2018).

Tabla 5.7

Resultados de la prueba post hoc basada en medias recortadas

Distancia entre medias aritméticas	Media recortada de confiabilidad al 10%		Ψ	95% IC	Significancia estadística
	Grupos	Población			
0.5	.860	.867	-.00672	[-.00695,- .00649]	<.001*
1.0	.877	.883	-.00646	[-.00667,- .00625]	<.001*
1.5	.897	.902	-.00520	[-.00537,- .00504]	<.001*
2.0	.915	.919	-.00413	[-.00426,- .00400]	<.001*

Nota. IC = intervalo de confianza. Los valores de Ψ no fueron redondeados para facilitar la interpretación de los intervalos de confianza.

*p<.05.

En la tabla 5.7 se presentan los resultados de las comparaciones múltiples realizadas a través del estadístico Ψ. En general, se puede afirmar que existen diferencias estadísticamente significativas en la confiabilidad estimada en cada uno de los cuatro escenarios simulados de distancia entre medias $\Psi_1 = -.00672,\ p < .001;\ \Psi_2 = -.00646,\ p < .001;\ \Psi_3 = -.00520,\ p < .001;\ \Psi_4 = -.00413,\ p < .001$).

Hasta ahora se ha reportado exclusivamente la significancia estadística de los resultados obtenidos; sin embargo, la American Statistical Association [ASA] declara que este indicador por sí solo no provee evidencia suficiente para realizar conclusiones de los resultados de investigaciones (Harvey y Brinkhof, 2019; Herbert, 2019; Schreiber, 2019; Staggs, 2019; Tarran, 2019; Wasserstein y Lazar, 2016; Wasserstein, Schirm y Lazar; 2019). Además, diversos autores critican que varias investigaciones utilizan el valor *p* como una especie de punto de corte que delimita la interpretación de los resultados de las investigaciones (Amrhein, Greenland y McShane; 2019; Andrade; 2019; Cassidy, Dimova, Giguère, Spence y Stanley, 2019; Wasserstein et al., 2019).

Por estos motivos, la tendencia actual sugiere abandonar la noción de significancia estadística y dar mayor énfasis a la significancia práctica del estudio, la cual se obtiene al estimar el tamaño del efecto (Andrade, 2019; Fritz, Morris y Richler, 2012; Pek y Flora, 2018; Wasserstein et al., 2019).

Este indicador permite cuantificar la magnitud de las diferencias encontradas en el experimento, lo que deriva en una mejor interpretación de los resultados (Sullivan y Feinn, 2012).

En este sentido, la estimación del tamaño del efecto puede realizarse a partir de diversas técnicas estadísticas estandarizadas.; no obstante, Leland Wilkinson y la Task Force on Statistical Inference (1999) afirman que, cuando las unidades de medición de una escala tienen un significado práctico, entonces es preferible utilizar un método no estandarizado para la estimación del tamaño del efecto. Entre estos métodos, el más sencillo y, sobre todo, *suficiente* es la simple diferencia entre las medias aritméticas (Fritz et al., 2012).

Tabla 5.8

Tamaño del efecto en función a la diferencia de medias de la confiabilidad estimada en cada escenario del experimento

Distancia entre medias de habilidad	*M*		Diferencia de medias
	Grupos	Población	
0.5	.860	.867	-.007
1.0	.877	.883	-.006
1.5	.897	.902	-.005
2.0	.915	.919	-.004

Aunque detectaron diferencias estadísticamente significativas al comparar la confiabilidad estimada en los cuatro escenarios de distancia entre medias aritméticas de habilidad, en la tabla 5.8 se puede observar que el tamaño del efecto es casi imperceptible en cada escenario. En otras palabras, las diferencias detectadas no tienen una implicancia práctica relevante, pues la magnitud de dichas diferencias es ínfima ($M_{g_1} - M_{p_1} = -.007$; $M_g - M_{p_2} = -.004$; $M_{g_3} - M_{p_3} = -.005$; $M_{g_4} - M_{p_4} = -.004$). La apreciación gráfica de este resultado se muestra en la figura 5.3.

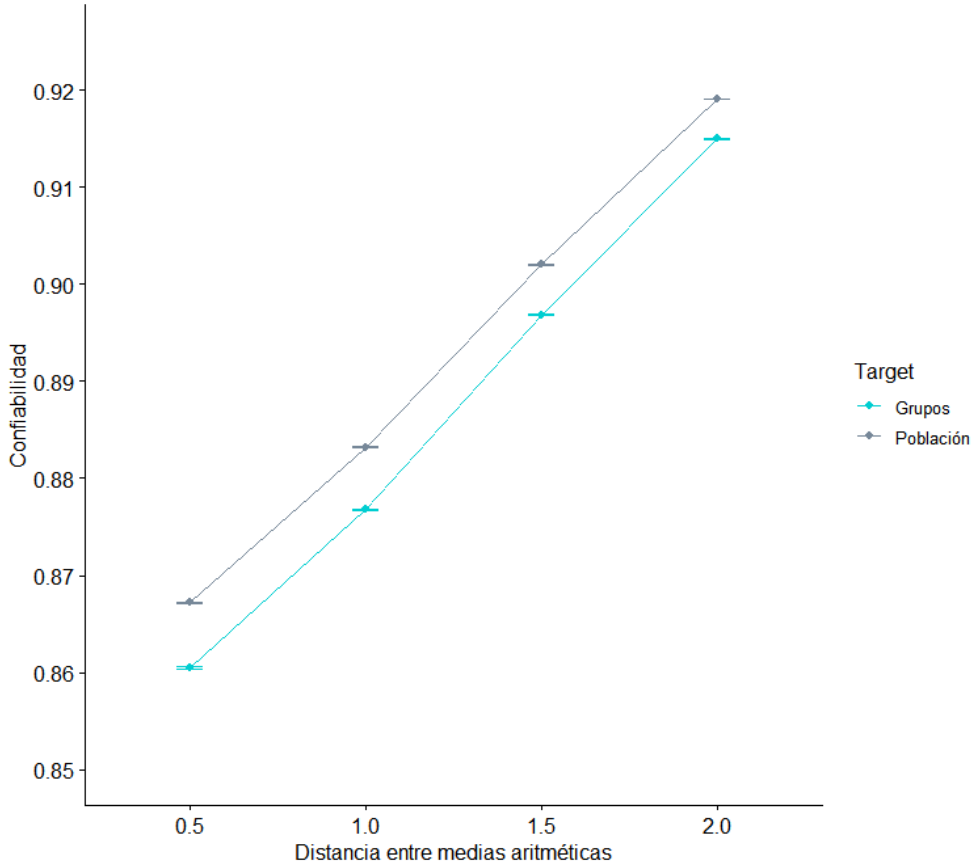

Figura 5.3. Gráfico lineal de las diferencias entre las medias aritméticas de la confiabilidad estimada en cada condición del experimento.

El código de programación desarrollado para realizar todos los análisis estadísticos y producción de gráficos se adjunta en el apéndice 2 con el objetivo de fomentar la reproducibilidad y replicabilidad de los resultados de este estudio, tal y como se recomienda en la literatura especializada (Stodden, Guo y Ma, 2013).

CAPÍTULO VI: DISCUSIÓN

La premisa que subyace al empleo de la estrategia del targeting establece que ensamblar tests seleccionando ítems en función al nivel de habilidad de las personas reducirá el error de estimación, brindando medidas más precisas (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Jones y Wright, 1992; Linacre, 2019a; Wright y Stone, 1979; 1999). Sobre la base de esta misma lógica, cuando los ítems se alejan del rango de habilidad de las personas, la confiabilidad de las puntuaciones se ve comprometida. Esto sucede porque la presencia de ítems off target implican situaciones en donde las personas con un alto nivel de habilidad puedan enfrentarse a ítems con muy poca dificultad y; del mismo modo, que personas con un nivel bajo de habilidad se enfrenten a ítems con alta dificultad. Dichas situaciones evocan conductas no deseables durante un proceso de evaluación como la falta de interés, adivinación o estilos de respuesta, e incrementan el error de estimación de parámetros (Ingebo, 1997; Wright y Stone, 1979).

Cuando se evalúa a toda una población a partir de un mismo test es posible que algunas personas se enfrenten a ítems off target y; por lo tanto, la confiabilidad de las puntuaciones se verá afectada. Por esta razón, la AERA, APA y NCME (2014) recomiendan la estimación del grado de confiabilidad en las puntuaciones obtenidas por cada grupo relevante de la población, pues las características particulares de cada uno de ellos pueden implicar una fuente importante de error de medición (Kubiszyn y Borich, 2013; Urbina, 2014).

Una estrategia para abordar dicha problemática consiste en el desarrollo de formas del test ensambladas empleando una estrategia de targeting en función al nivel de habilidad de cada grupo de la población (Boone et al., 2014). De acuerdo a las afirmaciones de diversos autores en el marco del modelo de medición Rasch (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Íngebo, 1997; Linacre, 2000; Wright y Stone, 1979; 1999), emplear formas diferenciadas del test para cada grupo poblacional utilizando la estrategia targeting debería resultar en mayor grado de confiabilidad de las puntuaciones.

Con el fin de evaluar esta premisa, se diseñó el presente estudio experimental en donde se compara la confiabilidad obtenida a partir de la aplicación de tests on target para toda la población con aquella obtenida al utilizar tests on target para dos grupos poblacionales en situaciones que varían de acuerdo a la distancia entre las medias aritméticas de las distribuciones de habilidad de ambos grupos.

En concordancia con la gran cantidad de autores en la literatura que afirman que la estrategia targeting mejora substancialmente el grado de precisión de las puntuaciones (Bond y Fox, 2015; Boone et al., 2014; Cavanagh y Waugh, 2011; Engelhard y Wind, 2018; Jones y Wright, 1992; Linacre, 2019a; Wright y Stone, 1979; 1999), se propuso como hipótesis de investigación que la confiabilidad estimada sería mayor al emplear pruebas diferenciadas on target para los grupos poblacionales. Además, conforme la diferencia entre las medias aritméticas del nivel de habilidad de los grupos que componen la población incrementa, la probabilidad de enfrentarse a ítems off target también y; por ello, la ganancia en confiabilidad a favor del targeting en grupos poblacionales será mayor. No obstante, los resultados obtenidos a partir del experimento Monte Carlo indican un efecto distinto al establecido como hipótesis.

El análisis de varianza factorial y las pruebas de comparaciones múltiples a través de métodos de estimación robustos denotan que existen diferencias estadísticamente significativas al comparar las medias aritméticas de la confiabilidad estimada a partir de un test enfocado en la población con aquella estimada a partir de test son target para cada grupo. Sin embargo, en contraste con la hipótesis planteada, el grado de confiabilidad estimado fue mayor al emplear un solo test on target para toda la población, en cada una de las situaciones de distancias de medias aritméticas simuladas.

Para una interpretación apropiada de estos resultados, es importante señalar que la American Statistical Association [ASA] afirma que muchos investigadores cometen el error de interpretar el valor *p* como un punto de corte que determina las conclusiones de los resultados de una investigación (Wasserstein et al., 2019). En este sentido, si se cometiera dicho error se afirmaría que emplear un targeting para la población resultaría en un mayor grado de confiabilidad de las puntuaciones, lo cual es una conclusión errónea. En su lugar, diversos autores recomiendan poner mayor énfasis en la significancia práctica de las diferencias detectadas a través de medidas del tamaño del efecto (Betensky, 2019; McShane, Gal, Robert y Tackett, 2019; Staggs, 2019).

Considerando dichas recomendaciones, se encontró que las diferencias detectadas a través de las pruebas de comparaciones múltiples tienen un tamaño del efecto ínfimo. Por lo tanto, se concluye que las diferencias en la confiabilidad estimada a partir del empleo de la estrategia targeting para toda la población y para los grupos no implican alguna significancia práctica.

La explicación de los resultados observados se encuentra estrechamente ligada a los factores que influyen en la estimación del índice de confiabilidad de separación de personas. De acuerdo a Linacre (2019a), dichos factores involucran el rango de habilidad de la muestra, la extensión del test, el número de categorías de respuesta por ítem y la estrategia targeting empleada. El control de cada uno de estos factores se estableció en el diseño del experimento; por ejemplo, se estableció el mismo rango de habilidad para la evaluación on target para dos grupos y aquella on target para toda población, pues los parámetros de habilidad de dicha población se componían a partir de la concatenación de los parámetros de los dos grupos; sin embargo, es relevante mencionar que esta acción asegura el mismo rango, pero no necesariamente la misma variabilidad de parámetros de habilidad.

Del mismo modo, se contempló una extensión equitativa de 40 ítems en cada forma del test; adicionalmente, todas las simulaciones se realizaron sobre la base del modelo Rasch para ítems dicotómicos; por lo tanto, todas las formas del test tuvieron el mismo formato de categorías de respuesta; y, finalmente, se empleó la misma metodología de targeting para ensamblar ambos tipos de evaluación. En consecuencia, la homogeneidad de las condiciones entre ambos tipos de evaluación podría ser una explicación de por qué no se encontraron diferencias substanciales entre la confiabilidad estimada a partir de ambas estrategias de targeting.

En otra instancia, los resultados del experimento también demuestran que, conforme aumenta la distancia entre las medias aritméticas de las distribuciones de habilidad de los grupos poblacionales, el grado de confiabilidad estimado incrementa. Este efecto puede explicarse en relación al incremento de la variabilidad en los datos; Linacre (2019a) afirma explícitamente que ante un mayor rango del nivel de habilidad de la muestra, mayor será el valor del índice de confiabilidad de separación de personas. En este sentido, conforme la distancia entre las medias aritméticas incrementa, el rango

habilidad de toda la población también lo hace y; en consecuencia, la confiabilidad de las puntuaciones es cada vez mayor.

Otro aspecto importante a considerar es la incompatibilidad entre la estrategia targeting y la equiparación de las medidas. Un adecuado proceso de equiparación en el contexto de los modelos TRI y Rasch consiste en que las formas distintas de un test compartan una serie de ítems que permitan realizar estimaciones de ambas escalas en una misma métrica (Linacre, 2019a). No obstante, es necesario que dichos ítems se extiendan en los distintos niveles de habilidad de la variable latente (Kolen y Brennan, 2014). En este sentido, se evidencia la incompatibilidad entre la estrategia targeting y la equiparación óptima de ítems, pues conforme la diferencia de habilidad entre dos grupos incrementa, el rango disponible para designar ítems en común se restringe y; por lo tanto, no es posible contar con ítems para cada nivel del rasgo latente en ambas formas del test, comprometiendo la estimación de parámetros.

En un experimento Monte Carlo análogo, DeMars (2002) evaluó el efecto de un diseño de equiparación entre dos grupos, en donde los ítems en común representan los ítems más difíciles de la forma entregada al grupo con menor habilidad y los más fáciles de la forma otorgada al grupo con mayor habilidad. Los resultados indican que este diseño resulta en la sobreestimación de los parámetros de los ítems únicos de la forma con mayor dificultad y en la subestimación de los parámetros de los ítems únicos de la forma con menor dificultad cuando se emplea el estimador MMLE. A pesar de que Gonzales y Wiberg (2017) argumenten que los resultados del estudio de simulación de DeMars (2002) no representan diferencias relevantes en la práctica, esta podría ser una explicación de por qué la confiabilidad obtenida a través del targeting en la población fue ínfimamente mayor a aquella obtenida al emplear un targeting en los grupos.

En conclusión, los resultados indican que emplear pruebas diferenciadas para grupos poblacionales no implica un incremento en la confiabilidad en las condiciones establecidas en la simulación. Tomando en cuenta que establecer formas diferenciadas de un test implica una serie de consideraciones en un proceso de evaluación a gran escala como en el diseño de nuevos ítems orientados a determinados niveles de dificultad (Bond y Fox, 2015), la evaluación piloto de dichos ítems (Kingston et al., 2013), las metodologías de ensamblaje de pruebas (Gonzales y Rutkowski, 2010), logística y recursos económicos asociados a la producción y administración de formas alternas de

un test (AERA et al, 2014), el análisis psicométrico de cada una de las formas (Hollenbeck, 2002), la divulgación de resultados fundamentando el empleo de formas diferenciadas (ITC, 2018), entre otros; no existe un argumento que sustente la implementación de esta estrategia en el diseño y construcción de instrumentos de evaluación a gran escala.

Como comentario final, es relevante reconocer que una limitación asociada a los experimentos Monte Carlo es el grado en que los resultados obtenidos puedan ser generalizados ante condiciones distintas a las designadas en el estudio (Feinberg y Rubright, 2016). En este sentido, los resultados se limitan específicamente a las condiciones establecidas y; por ello, alguna modificación en alguno de los factores controlados tales como en la variabilidad de niveles de habilidad, categorías de respuesta, extensión de pruebas, tamaño de muestra, método de estimación, modelo de medición, diseño de equiparación o incluso el tipo de estrategia targeting que se desee emplear puede implicar un resultado distinto al obtenido.

CONCLUSIONES

- Existe un efecto estadísticamente significativo de la distancia entre medias aritméticas de dos grupos que componen una población, el target de la evaluación y la interacción entre ambos factores en la confiabilidad de las puntuaciones.
- Existen diferencias estadísticamente significativas al comparar la confiabilidad estimada a partir de un test on target para toda la población y aquella obtenida al emplear pruebas diferenciadas on target para dos grupos poblacionales en todos los escenarios de diferencias entre medias aritméticas simulados; sin embargo, dichas diferencias son irrelevantes en la práctica.
- Conforme aumenta la diferencia entre las medias aritméticas del nivel de habilidad de los grupos poblacionales incrementa el grado de confiabilidad de las puntuaciones debido al aumento en la variabilidad de los niveles de habilidad.
- La estrategia de targeting es incompatible con el proceso de equiparación de los ítems en el marco del modelo de medición Rasch, pues conforme aumenta la diferencia entre las medias aritméticas de habilidad de las personas, se reduce el rango disponible para designar ítems en común.
- Los resultados de esta investigación demuestran que no existe un fundamento sólido para optar por tests on target para cada grupo relevante de la población, específicamente en las condiciones simuladas.

RECOMENDACIONES

- Optar por el diseño y construcción de instrumentos de evaluación a gran escala empleando la estrategia tradicional del targeting enfocado en el nivel de habilidad de toda la población. Al no existir diferencias entre ambas estrategias de ensamblaje de ítems, se sugiere emplear el target en toda la población porque involucra menor complejidad en cuestión a logística, metodología y recursos.
- Si se desea implementar formas diferenciadas de una evaluación empleando tests on target para cada grupo, considerar que la diferencia entre las medias aritméticas puede reducir el rango disponible para designar ítems en común; por lo tanto, es posible que afecte al proceso de equiparación de puntuaciones.
- Finalmente, se recomienda replicar el experimento considerando la modificación de distintas condiciones de evaluación como el tamaño de muestra de cada grupo, longitud de los tests, número de categorías de respuesta, diseño de equiparación, modelo de medición, distribución de parámetros, y métodos de estimación; con el objetivo de ampliar el grado de generalización de los resultados del experimento a otros escenarios.

REFERENCIAS

Almond, P. J., Lehr, C., Thurlow, M. L., & Quenemoen, R. (2002). Participation in large-scale state assessment and accountability systems. En T. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 314-370). Mahwah, NJ: Lawrece Erlbaum.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Amrhein, V., Greenland, S., McShane, B. (2019). Scientists rise up against statistical significance. *Nature 567*(7748), 305-307. https://doi.org/10.1038/d41586-019-00857-9

Andrade, C. (2019). The *p* value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, *41*(3), 210-215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: SAGE Publications, Inc.

Andrich, D. (2004). Controversy and the Rasch Model. *Medical Care*, *42*(1), 7-16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory. Measuring in the educational, social and health sciences*. Singapore: Springer.

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3), 1105-1106.

Balakrishnan, N., Koutras, M. V., & Politis, K. G. (2020). *Introduction to probability models and applications*. Hoboken, NJ: John Wiley & Sons, Inc.

Barker, C., Pistrang, N., & Elliott, R. (2016). *Research methods in clinical psychology: An introduction for students and practitioners* (3rd ed.). Chichester: John Wiley & Sons, Inc.

Bartram, D., & Amado, N. (2017). Psychological assessment, standards and guidelines for. *Reference Module in Neuroscience and Biobehavioral Psychology*, 1-5. https://doi.org/10.1016/B978-0-12-809324-5.05681-9

Bartram, D., & Hambleton, R. K. (2016). The ITC guidelines: International standards and guidelines relating to tests and testing. En F. T. Leong, D., Bartram, F. M., Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 35-48). New York, NY: Oxford University Press.

Baur, T., & Lukes, D. (2009). An evaluation of the IRT models through Monte Carlo simulation. *University of Wisconsin-La Crosse Journal of Undergraduate Research XII*, 1-7.

Beiser, D., Vu, M., & Gibbons, R. (2016). Test-retest reliability of a computerized adaptive depression test. *Psychiatric Services, 67*(9), 1039-1041.

Berezner, A., & Adams, R. J. (2017). Why large-scale assessments use scaling and Item Response Theory. En P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 323-356). Chichester: John Wiley and Sons, Inc.

Betensky, R. A. (2019). The p-value requires context, not a threshold. *The American Statistician, 73*(1), 115-117. https://doi.org/10.1080/00031305.2018.1529624

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/bf02293801

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*(2), 179–197. https://doi.org/10.1007/bf02291262

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444. https://doi.org/10.1177/014662168200600405

Bond, T. G., & Fox, C. M. (2015*). Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE-Life Sciences Education, 15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer.

Bordens, K. S., & Abbott, B. B. (2018). *Research design and methods* (3rd ed.). New York, NY: McGraw Hill.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues on contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D., & Scholten, A. Z. (2008). The Rasch model and Conjoint Measurement theory from the perspective of psychometrics. *Theory & Psychology, 18*(1), 111–117. https://doi.org/10.1177/0959354307086925

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. Psychological Review, 110(2), 203-219. https://doi.org/10.1037/0033-295X.110.2.20

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*(1), 5–20. https://doi.org/10.1177/0013164406288162

Breithaupt, K., & Hare, D. R. (2016). Automated test assembly. En F. Drasgow (Ed.), *Technology in testing* (pp. 128-141). New York, NY: Routledge.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-317.

Bridgman, P. W. (1927). *The Logic of Modern Physics*. New York, NY: The MacMillan Company.

Brough, P., & Hawkes, A. (2019). *Designing impactful research.* En P. Brough (Ed.), *Advanced research methods for applied psychology: Design, analysis and reporting* (pp. 7-14). New York, NY: Routledge.

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 266-287. https://doi.org/10.21031/epod.305821

Callingham, R., & Bond, T. G. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal, 18*(2), 1-10. https://doi.org/10.1007/BF03217432

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293. https://doi.org/10.2307/1165298

Campbell, N. R. (1920). *Physics. The elements*. London: Cambridge University Press.

Campbell, N. R. (1921). *What is Science?* London: Methuen & Co. Ltd

Carey, S. S. (2004). *A beginner's guide to scientific method* (4th ed.). United Kingdom: Wadsworth Cengage Learning.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019) Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science.* Advance online publication. https://doi.org/10.1177/2515245919858072

Cavanagh, R., & Waugh, R. F. (2011). *Applications of Rasch measurement in learning environments research*. Rotterdam: Sense Publishers.

Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of Maximum Likelihood Estimation and Expected a Posteriori Estimation in CAT using the Partial Credit Model. *Educational and Psychological Measurement, 58*(4), 569–595. https://doi.org/10.1177/0013164498058004002

Chen, S. K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the Rating Scale Model. *Educational and Psychological Measurement, 57*(3), 422–439. https://doi.org/10.1177/0013164497057003004

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch models in health.* Hoboken, NJ: ISTE Ltd & John Wiley & Sons, Inc.

Ciccarelli, S. K., & White, J. N. (2018). *Psychology* (5th ed.). Harlow: Pearson

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186-190.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. Thousand Oaks, CA: SAGE Publications, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). Nueva York: McGraw Hill

Coolican, H. (2014). *Research methods and statistics in psychology* (6th ed.). New York, NY: Psychology Press.

Coolican, H., & Kelly, O. (2014). *Research methods in psychology*. Ontario: Oxford University Press.

Cooper, C. (2019). *Psychological testing: Theory and practice*. New York, NY: Routledge.

Coulacoglou, C., & Saklofske, D. (2017). *Psychometrics and psychological assessment*. London: Academic Press.

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. Washington, DC: OECD Publishing. https://doi.org/10.1787/9789264248373-en

Creswell, J. W., & Creswell, J. D. (2018*). Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: John Wiley & Sons, Inc.

Custer, M. (2015, October 21-24). *Sample size and ítem parameter estimation precisión when utilizing the one-parameter "Rasch" model*. Paper presented at the Annual Meeting of The Mid-Western Educational Research Association, Evanston, IL.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education, 15*(1), 15–31. https://doi.org/10.1207/s15324818ame1501_02

Denis, D. J. (2019). *SPSS data analysis for univariate, bivariate, and multivariate statistics*. New York, NY: John Wiley & Sons, Inc.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. New York, NY: CRC Press.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science, 1*, 5–26. https://doi.org/10.1093/bjps/I.1.5

Drasgow, F. (1989). An evaluation of Marginal Maximum Likelihood Estimation for the Two-Parameter Logistic Model. *Applied Psychological Measurement, 13*(1), 77-90. https://doi.org/10.1177/014662168901300108

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequeal variance case. *Journal of the American Statistical Association, 75*, 796-800. https://doi.org/10.2307/2287161

Edmonds, W. A., & Kennedy, T. D. (2017). *An applied guide to research designs. Quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Recuperado de: https://www.ets.org/s/about/pdf/standards.pdf

Ember, C. R., & Ember, M. (2009). *Cross-cultural research methods* (2nd ed.). Lanham: AltaMira Press.

Engelhard, G., & Wind, S. (2018). *Invariant measurement with raters and rating scales*. New York, NY: Routledge.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. https://doi.org/10.1111/emip.12111

Ferrer, J. G., y Arregui, P. (2002). *La experiencia latinoamericana con pruebas internacionales de aprendizaje: Impacto sobre los procesos de mejoramiento de la calidad de la educación y criterios para guiar las decisiones sobre nuevas aplicaciones*. Recuperado de: http://repositorio.minedu.gob.pe/handle/123456789/216.

Field, A. P. (2007). Analysis of variance. En N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 1, pp. 32-35). Thousand Oaks, CA: SAGE Publications, Inc.

Field, A. P. (2018). *Discovering statistics using IBM SPSS Statistics*. London: SAGE Publications, Inc.

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: SAGE Publications, Inc.

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. New York, NY: Routledge.

Fisher, W. P. (1994). The Rasch debate: Validity and revolution in educational measurement. En M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 36–72). Norwood, NJ: Ablex.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Fritz, C., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology General, 141*(1), 2-18. https://doi.org/10.1037/a0024338

Garson, G. D. (2013). *Validity & reliability*. Asheboro, NC: Statistical Associates Publishing

Geisinger, K. F. (2013). Reliability. En K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Volume 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 21-42) Washington, DC: American Psychological Association.

Ghasemi, A., & Zahedias, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology & Metabolism, 10*(2), 486-489. https://doi.org/10.5812/ijem.3505

Goldman, S. H.., & Raju, N. S. (1986). Recovery of one and two parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement, 46*(1), 11-21. https://doi.org/10.1177/0013164486461002

Gonzales, E. J., & Rutkowski, L. (2010). *Principles of multiples matrix booklet designs and parameter recovery in large-scale assessments*. Recuperado de http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf

Gonzales, J., & Wiberg, M. (2017). *Applying test equating methods using R* [E-reader version]. Springer.

Grami, A. (2020). *Probability, random variables, statistics, and random processes. Fundamentals & applications*. Hoboken, NJ: John Wiley & Sons, Inc.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altaman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337-350. https://doi.org/10.1007/s10654-016-0149-3

Guyer, R., & Thompson, N. (2011). *Item Response Theory parameter recovery using XcalibreTM 4.1*. Santi Paul, MN: Assessment Systems Corporation.

Haertel, E. H. (2006). Reliability. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Wesport, CT: American Council on Education y Praeger Publishers.

Hambleton, R. K., & Swaminathan, H. (1991). *Item Response Theory. Principles and applications*. New York, NY: Springer.

Harrel, F. E. (2015). *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Switzerland: Springer International.

Harrison, R. L. (2010). Introduction to Monte Carlo simulation. *AIP Conference Proceedings, 1204*(1), 17-21. https://doi.org/10.1063/1.3295638

Harvey, L. A., & Brinkhof, M. W. G. (2019). Imagine a research world without the words "statistically significant". Is it really possible? *Spinal cord, 57*, 437-438. https://doi.org/10.1038/s41393-019-0292-2

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in Item Response Theory. *Educational and Psychological Measurement, 57*(2), 266–279. https://doi.org/10.1177/0013164497057002006

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology, 4*, 1-4. https://doi.org/10.3389/fpsyg.2013.00246

Heise, D. R. (2015). Scaling and classification in social measurement. En J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 21, pp. 4-8). Amsterdam: Elsevier.

Herbert, R. (2019). Research note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. *Journal of Physiotherapy 65*(3), 178-181. https://doi.org/10.1016/j.jphys.2019.05.001

Heumann, C., Schomaker, M. (2016). *Introduction to statistics and data analysis. With exercises, solutions and applications in R* [E-reader version]. Springer.

Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. En L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 11-36). New York: CRC Press.

Hilmer, C. (2010). Bootstrapping. En N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 101-104). Thousand Oaks, CA: SAGE Publications, Inc.

Ho, R. (2018). *Understanding statistics for the social sciences with IBM SPSS*. Boca Raton, FL: CRC Press.

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*, 1-9. https://doi.org/10.3389/fpsyg.2012.00137

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. En G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 53-68). New York, NY: Srpinger-Verlag.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [Los axiomas de cantidad y la teoría de medición]. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse, 53*, 1–46.

Hollenbeck, K. (2002). Determining when test alterations are valid accomodations or modifications for large-scale assessment. En G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 395-426). New Jersey, NJ: Lawrence Erlbaum.

Horgan, J. M. (2020). (*Probability with R. An introduction with computer science applications* (2nd ed). Hoboken, NJ: John Wiley & Sons, Inc.

Howitt, D., & Cramer, D. (2017). *Research methods in psychology* (5th ed.). Slovakia: Pearson.

Humphry, S. (2013). Understanding measurement in light of its origins. *Frontiers in Psychology, 4*, 1-8. https://doi.org/10.3389/fpsyg.2013.00113

Ingebo, G. (1997). Linking tests with the Rasch model. *Rasch Measurement Transactions, 11*(1), 549.

International Test Comission. (2013a). *ITC guidelines on quality control in scoring, test analysis, and reporting of test scores*. Recuperado de: https://www.intestcom.org/files/guideline_quality_control.pdf

International Test Comission. (2013b). *ITC guidelines on test use*. Recuperado de: https://www.intestcom.org/files/guideline_test_use.pdf

International Test Comission. (2018). *ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations*. Recuperado de: https://www.intestcom.org/files/guideline_diverse_populations.pdf

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika, 42*(4), 567–578. https://doi.org/10.1007/bf02295979

Johnson, B., & Christensen, L. (2016). *Educational research. Quantitative, qualitative, and mixed approaches* (6th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. En M. O. Martin, & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* (pp. 1-21). Recuperado de https://timssandpirls.bc.edu/methods/

Jones, N., & Wright, B. D. (1992). Trials with vertical equating. *Rasch Measurement Transactions, 6*(3), 240.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 99–127. https://doi.org/10.1142/9789814417358_0006

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing. Principles, applications and issues* (9th ed.). Boston, MA: Cengage Learning.

Karabatsos, G. (2017). On bayesian testing of Additive Conjoint Measurement axioms using synthetic likelihood. *Psychometrika, 83*(2), 321–332. https://doi.org/10.1007/s11336-017-9581-x

Kelley, K. (2010). Monte Carlo simulation. En N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 831-832). Thousand Oaks, CA: SAGE Publications, Inc.

Khan, M. I. (2014). Recovery and stability of item parameter and model fit across varying sample sizes and test lengths in Rasch analysis with small sample. *Social Science International, 30*(1), 43-60.

Kingston, N. M., Scheuring, S. T., & Kramer, L. B. (2013). Test development strategies. En K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Volume 1. Test theory and testing and assessment in industrial and organizational*

psychology (pp. 165-184). Washington, DC: American Psychological Association.

Kirsch, I., Lennon, M., von Davier, M., Gonzales, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. En M. von Davier, E. Gonzales, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1-12). New York, NY: Springer.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression. A self-learning text* (3rd ed.). New York, NY: Springer.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. En M. von Davier, E. Gonzales, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115-148). New York, NY: Springer.

Kline, P. (2015). *A handbook of test construction*. New York, NY: Routledge.

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: Guilford Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Kosso, P. (2011). *A summary of scientific method.* New York: Springer.

Kruyen, P. (2012). *Using short tests and questionnaires for making decisions about individuals: When is short too short?* (doctoral dissertation). Recuperado de https://pure.uvt.nl/ws/portalfiles/portal/1467991/Kruyen_using_14-12-2012.pdf

Kubiszyn, T., & Borich, G. D. (2013). *Educational testing & measurement* (10th ed.). Hoboken. NJ: John Wiley & Sons, Inc.

Kyngdon, A. (2008). Conjoint Measurement, error and the Rasch model. *Theory & Psychology, 18*(1), 125–131. https://doi.org/10.1177/0959354307086927

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación. (2015). *Informe de resultados TERCE: Logros de aprendizaje.* Recuperado de: http://unesdoc.unesco.org/images/0024/002435/243532S.pdf

LaRoche, S., Joncas, M., & Foy, P. (2017). Sample design in PIRLS 2016. En M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 1-34). Recuperado de https://timssandpirls.bc.edu/ publications/pirls/2016-methods.html

Leppink, J. (2019). *Statistical methods for experimental research in education and psychology*. Switzerland: Springer.

Li, Y. (2010). Root mean square error. En N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1287-1288). Thousand Oaks, CA: SAGE Publications, Inc.

Lietz, P., & Tobin, M. (2016). The impact of large-scale assessments in education on education policy: evidence from around the world. *Research Papers in Education, 31*(5), 499–501. https://doi.org/10.1080/02671522.2016.1225918

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating Item Response Theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. https://doi.org/10.1037/0021-9010.75.2.164

Linacre J.M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions, 19*(3), 1032.

Linacre, J. M. & Wright, B. D. (1994). Chi – square fit statistics. En J. M. Linacre (Ed.), *Rasch Measurement Transactions Part 1* (pp. 360-361). Chicago: MESA Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1996). Sample size again. *Rasch Measurement Transactions, 9*(4), 468.

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement, 3*, 381-405.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come*. Recuperado de https://www.rasch.org/memo69.pdf

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2011). Rasch measures and unidimensionality. *Rasch Measurement Transactions, 24*(4), 1310.

Linacre, J. M. (2014, November 11). Re: Targeting between item difficulty and person ability [Online forum comment]. Recuperado de http://raschforum.boards.net/thread/173/targeting-item-difficulty-person-ability.

Linacre, J. M. (2019a). *A user's guide to WINSTEPS*. Recuperado de https://www.winsteps.com/a/Winsteps-Manual.pdf

Linacre, J. M. (2019b, April 29). A question about targeting [Online forum comment]. Recuperado de http://raschforum.boards.net/thread/1098/question-targeting.

Loe, B. S., Stillwell, D., & Gibbons, C. (2017). Computerized adaptive testing provides reliable and efficient depression measurement using the CES-D scale. *Journal of Medical Internet Research, 19*(9), 1-13. https://doi.org/10.2196/jmir.7453.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51-61). New York, NY: Academic Press.

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Luecht, R. M. (2014). Computer-based and computer-adaptive testing. En M. Simon, K. Ercikan, & M. rousseau (Eds.), *Improving large-scale assessment in education. Theory, issues, and practice* (pp. 62-82). New York, NY: Routledge.

Lunz, M. E. (2010). *Using the very useful Wright map*. Recuperado de: https://www.rasch.org/mra/mra-01-10.htm

Lunz, M. R. (2009). *Test length and test reliability for multiple choice examinations*. Recuperado de: https://www.rasch.org/mra/mra-02-09.htm.

Mair, P. (2018). *Modern psychometrics using R* [E-reader version]. Springer. https://doi.org/10.1007/978-3-319-93177-7

Mair, P., & Wilcox, R. (2019a). A collection of robust statistical methods [R Package] Recuperado de https://cran.r-project.org/web/packages/WRS2/WRS2.pdf

Mair, P., & Wilcox, R. (2019b). Robust statistical methods in R using the WRS2 package. *Behavioral Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-019-01246-w

Mallison, T., & Stelmack, J. (2001). Going beyond unreliable reliabilities. *Rasch Measurement Transactions, 14*(4), 787-788.

Marczyk, G., de Matteo, D., & Festinger, D. (2005). *Essential of research design and methodology*. Hoboken, NJ: John Wiley & Sons, Inc.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51*, 315–327. https://doi.org/10.1016/j.measurement.2014.02.014

Mari, L., Maul, A., Irribarra, D. T., & Wilson, M. (2016). A meta-structural understanding of measurement. *Journal of Physics: Conference Series, 772*, 1-6. https://doi.org/10.1088/1742-6596/772/1/012009

Mari, L., Maul, A., Torres-Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100,* 115–121. https://doi.org/10.1016/j.measurement.2016.12.050

Matthews, B., & Ross, L. (2010). *Research methods. A practical guide for the social sciences*. Harlow: Pearson Education Limited.

Maul, A., Torres-Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79,* 311-320. https://doi.org/10.1016/j.measurement.2015.11.001

Maul, A., Wilson, M., & Torres-Irribarra, D. (2013). On the conceptual foundations of psychological measurement. *Journal of Physics: Conference Series, 459,* 1-6. https://doi.org/10.1088/1742-6596/459/1/012008

Maxwell, J. C. (1873). *A Treatise on Electricity and Magnetism*. Oxford: Clarendon Press.

McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGrane, J. A. (2010). Are psychological quantities and measurement relevant in the 21st century? *Frontiers in Psychology, 1,* 1-2. https://doi.org/10.3389/fpsyg.2010.00022

McGrane, J. A. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology, 6,* 1-8. https://doi.org/10.3389/fpsyg.2015.00431

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician, 73*(1), 235-245. https://doi.org/10.1080/00031305.2018.1527253

Mendelovits, J. (2017). Test development. En P. Lietz, J. C. Cresswell, K. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 63-90). Chichester: John Wiley & Sons, Inc.

Michell, J. (1990/2014). *An introduction to the logic of psychological measurement*. New York, NY: Psychology Press.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*(5), 639-667. https://doi.org/ 10.1177/0959354300105004

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*, 211-217. https://doi.org/10.1080/00050060108259657

Michell, J. (2002). Steven's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54*, 99-194. https://doi.org/ 10.1080/00049530210001706563

Michell, J. (2007). Measurement. En S. P. Turner, & M. W. Risjord (Eds.), *Philosophy of anthropology and sociology* (pp. 71-120). Netherlands: Elsevier.

Michell, J. (2008). Conjoint measurement and the Rasch paradox. *Theory & Psychology, 18*(1), 119-124. https://doi.org/ 10.1177/0959354307086926

Michell, J. (2009). Invalidity in Validity. En R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 111-133). Charlotte, NC: Information Age Publishing.

Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma. *Theory & Psychology, 24*(1), 11-123. https://doi.org/10.1177/0959354313517524

Michell, J. (2019). Conjoint measurement underdone: Comment on Günter Trendler (2019). *Theory & Psychology, 29*(1), 138-143. https://doi.org/ 10.1177/0959354318814962

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement: Translated from part I of Otto Hölder's german text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology, 40*(3), 235-252. https://doi.org/10.1006/jmps.1996.0023

Miller, L. (2013). *Cross-cultural research with integrity*. New York, NY: Palgrave Macmillan.

Ministerio de Educación. (2014). *Reporte técnico de la evaluación muestral 2013.* Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2016/08/archivo_web.pdf

Ministerio de Educación. (2018). *Reporte técnico de la evaluación censal de estudiantes (ECE 2016).* Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2018/03/Reporte-Tecnico-ECE-2016.pdf

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1), 67-72.

Mislevy, R. J., & Stocking, M. L. (1989). A Consumer's Guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*(1), 57–75. https://doi.org/10.1177/014662168901300106

Morling, B. (2015). *Research methods in psychology* (3th ed.). New York, NY: W. W. Norton & Company, Inc.

Mundform, D. J., Schaffer, J., Kim, M. J., Shaw, D., Thongteeraparp, A., & Supawan, P. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods, 10*(1), 19-28.

Nagel, E., & Hempel, C. G. (1931). Measurement. *Erkenntnis, 2*(1), 313–335. https://doi.org/10.1007/bf02028166

Olvera, O. S. (2017). *On Mone Carlo simulation algorithms for research in psychometrics* (doctoral dissertation). The University of British Columbia, Vancouver, Canadá.

Organization for Economic Co-operation and Development. (2016). *Sampling in PISA*. Recuperado de: https://www.oecd.org/pisa/pisaproducts/SAMPLING-IN-PISA.pdf

Organization for Economic Co-operation and Development. (2017). *PISA 2015 technical report.* Recuperado de: https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Osborne, J. W. (2010). Challenges for quantitative psychology and measurement in the 21st century. *Frontiers in Psychology, 1*, 1–3. https://doi.org/10.3389/fpsyg.2010.00001

Paek, I., & Cole, K. (2020). *Using R for Item Response Theory model applications*. New York, NY: Routledge.

Paolella, M. S. (2019). *Linear models and time-series analysis. Regression, ANOCA, ARMA and GRACH*. Oxford: John Wiley and Sons, Inc.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287-312. https://doi.org/10.1207/S15328007SEM0802_7

Pearson, R. K. (2018). *Exploratory data analysis using R*. Boca Raton, FL: CRC Press.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208-225. https://doi.org/10.1037/met0000126

Penfield, R. D. (2013). Item analysis. En K. F. Geisinger (Ed.), *APA Handbook of testing and assessment in psychology. Volume 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 121-138). Washington, DC: American Psychological Association. Associates.

Peng, Ch-Y., J., & Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3-14. https://doi.org/10.1080/00220670209598786

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as Additive Conjoint Measurement. *Applied Psychological Measurement, 3*(2), 237–255. https://doi.org/10.1177/014662167900300213

Pohlert, T. (2019). *Calculate pairwise multiple comparisons of mean rank sums extended* [R package]. Recuperado de https://cran.r-project.org/ web/packages/ PMCMRplus/PMCMRplus.pdf

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. Vienna: R Foundation for Statistical Computing.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement, 21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Reise, S. P, Moore, T. M. & Haviland, M. G. (2013). Applying unidimensional item response theory models to psychological data. En K. F. Geisinger (Ed.), *APA Handbook of testing and assessment in psychology. Volume 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 101-120). Washington, DC: American Psychological Association.

Resse, T. W. (1943). The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples. *Psychological Monographs, 55*(3), 1–89. https://doi.org/10.1037/h0093539

Revelle, W. (2019). *Procedures for psychological, psychometric, and personality research* [R Package]. Recuperado de https://cran.r-project.org/web/packages/psych/psych.pdf

Revelle, W. (2019). *Using R and the psych package to find ω*. Recuperado de http://personality-project.org/r/psych/HowTo/omega.pdf

Rizopoulos, D. (2018). *Latent trait models under IRT* [R Package]. Recuperado de https://cran.r-project.org/web/packages/ltm/ltm.pdf

Robitzsch, A., Keifer, T., & Wu, M. (2019). *Test analysis modules* [R Package]. Recuperado de https://cran.r-project.org/web/packages/TAM/TAM.pdf

Rogers, H. J., & Swaminathan, H. (2016). Concepts and methods in research on Differential Functioning of test items. Past, present and future. En C. S. Wells, & M. Faulkner-Bond (Eds.), *Educational measurement. From foundations to future* (pp. 126-142). New York, NY: The Guilford Press.

Roni, S. M., Merga, M. K., & Morris, J. E. (2020). *Conducting quantitative research in education*. New York, NY: Springer.

Russell, B. (1903/2010). *Principles of mathematics*. New York, NY: Routledge.

Rutherford, A. (2011). *ANOVA and ANCOVA. A GLM approach* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Sahin, A., & Aml, D. (2016). The effects of test length and simple size on item parameters in Item Response Theory. *Educational Sciences: Theory and Practice, 17*, 321-335. https://doi.org/10.12738/estp.2017.1.0270

Salkind, N. J. (2018). *Exploring research* (9th ed.). Harlow: Pearson.

Schaw, F. (1991). Descriptive IRT vs. Prescriptive Rasch. *Rasch Measurement Transactions, 5*(1), 131.

Schreiber, J. B. (2019). New paradigms for considering statistical significance: A way forward for health services research journals, their authors, and their readership.

Research in Social and Administrative Pharmacy. Advance online publication. https://doi.org/10.1016/j.sapharm.2019.05.023

Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394–409. https://doi.org/10.1177/0013164406294776

Seong, T. J. (1990). Sensitivity of Marginal Maximum Likelihood Estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299–311. https://doi.org/10.1177/014662169001400307

Sijtsma, K. (2008). Reliability beyond theory and into practice. *Psychometrika*, *74*(1), 169–173. https://doi.org/10.1007/s11336-008-9103-y

Sireci, S. G., & Gándara, F. (2016). Testing in educational and developmental settings. En F. T. L. Leong, D. Bartram, F. M., Cheung, K. F., Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 187-202). New York, NY: Oxford University Press.

Smith, R. M. (1991). Guessing and the Rasch model. *Rasch Measurement Transactions, 6*(4), 262-263.

Staggs, V. S. (2019). Why statisticians are abandoning statistical significance. *Research in Nursing & Health 42*(3), 159-160. https://doi.org/10.1002/nur.21947

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680. https://doi.org/10.1126/science.103.2684.677

Stodden, V., Guo, P., Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE 8*(6), 1-8. https://doi.org/10.1371/journal.pone.0067111.

Stone, M., & Yumoto, F. (2004). The effect of sample size on Rasch/IRT parameters using dichotomous items. *Journal of Applied Measurement, 5*(1), 48-61.

Streiner, D. L. (2010). Measure for measure: New developments in measurement and Item Response Theory. *The Canadian Journal of Psychiatry, 55*(3), 180-186. https://doi.org/10.1177/070674371005500310

Sullivan, G. M., & Feinn, R. (2012). Using effect size – or why the *p* value is not enough. *Journal of Graduate Medical Education, 4*(3), 279-282. https://doi.org/10.4300/JGME-D-12-00156.1

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., …, Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling, 55*(4), 335-360

Tarran, B. (2019). Is this the end of "statistical significance"? *Significance 16*(2), 4-5. https://doi.org/10.1111/j.1740-9713.2019.01244.x

Tendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19*, 579–599.

Tennant, A., & Pallant, J. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions, 20*, 1082-1084.

Ter Laak, J. F., Gokhale, M., & Desai, D. (2013). *Understanding psychological assessment: A primer on the global assessment of client's behavior in educational and organizational setting*. India: SAGE Publications, Inc.

The British Psychological Society. (2017). *Psychological testing: A test user's guide*. Recuperado de https://ptc.bps.org.uk/information-and-resources/information-testing/guidelines-testing-and-test-use.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. En G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 1-26). New Jersey, NJ: Lawrence Erlbaum.

Tinsley, H. E. A., & Brown, S. D. (2000). *Handbook of applied multivariate statistics and mathematical modeling*. London: Academic Press.

Tobin, M., Nugroho, D., & Lietz, P. (2016). Large-scale assessments of students' learning and education policy: Synthesising evidence across world regions. *Research Papers in Education, 31*(5), 578–594. https://doi.org/10.1080/02671522.2016.1225353

Tuckey, J. W. (1977). *Exploratory data analysis*. London: Addison-Wesley Publishing Company

United Nations Educational, Scientific and Cultural Organization & Oficina Regional de Educación para América Latina y el Caribe. (2016). *Informe de resultados TERCE: Logros de aprendizaje*. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000243532

United Nations Educational, Scientific and Cultural Organization & Oficina Regional de Educación para América Latina y el Caribe. (2009). *Reporte técnico SERCE: Los aprendizajes de los estudiantes de América Latina y el Caribe*. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000190297?posInSet=3&queryId=1672db84-8717-4afe-9826-5f6384b16924

United Nations Educational, Scientific and Cultural Organization. (2018). *The impact of large-scale learning assessments*. Recuperado de http://uis.unesco.org/sites/default/files/documents/impact-large-scale-assessments-2018-en.pdf

United Nations Educational, Scientific and Cultural Organization. (2019). *The promise of large-scale learning assessments*. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000369697

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.), Hoboken, NJ: John Wiley & Sons, Inc.

Van der Linden. W. J. (2016). Introduction. En W. J. van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 1-12). New York, NY: CRC Press.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature, 514*, 550–553. https://doi.org/10.1038/514550a

von Davier, M. (2016). Rasch model. En W. J. van der Linden (Ed.), *Handbook of Item Response* (Vol. 1, pp.31-50). New York, NY: CRC Press.

Wagemaker, H. (2014). International large-scale assessment: From research to policy. En L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 11-36). New York: CRC Press.

Wahed, A. S., & Tang, X. (2010). Analysis of variance. En N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 26-29). Thousand Oaks, CA: SAGE Publications, Inc.

Walker, M. (2017). Computer-based delivery of cognitive assessment and questionnaires. En P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 231-252). Chichester: John Wiley & Sons, Inc.

Wang, S., & Wang, T. (2001). Precision of Warm's Weighted Likelihood Estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317–331. https://doi.org/10.1177/01466210122032163

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p<0.05". *The American Statistician, 73*(1), 1-19. https://doi.org/10.1080/00031305.2019.1583913

Waugh, R. F. (2007). Rasch measurement model. En N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 3, pp. 820-825). Thousand Oaks, CA: SAGE Publications, Inc.

Wilcox, R. (2010). *Fundamentals of modern statistical methods. Substantially improving power and accuracy* (2nd ed.). New York, NY: Springer.

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Ney York, NY: Elsevier.

Wilcox, R. (2017a). *Modern statistics for the social and behavioral sciences. A practical introduction* (2nd ed.). London: CRC Press.

Wilcox, R. (2017b). *Understanding and applying basic statistical methods using R*. Hoboken, NJ: John Wiley & Sons, Inc.

Wiley, M., & Wiley, J. F. (2019). *Advanced R statistical programming and data models. Analysis, machine learning, and visualization*. Columbia, IN: Apress.

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14*(3), 219–225. doi:10.1111/j.1745-3984.1977.tb00039.x

Wright, B. D. (1985). Additivity in psychological measurement. En E. Roskam (Ed.), *Measurement and personality assessment* (pp. 101–112). North Holland: Elsevier Science.

Wright, B. D. (1992). What is the "right" test length? *Rasch Measurement Transactions, 6*(1), 205.

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing*. Chicago, IL: MESA Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Jones, N. (1992). Trials with vertical equating. *Rasch Measurement Transactions, 6*(3), 240.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857-860.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range Inc.

Yang, Y., & Green, S. B. (2011). Coefficient Alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377–392. https://doi.org/10.1177/0734282911406668

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Pyschometrika, 52*(2), 275-291. https://doi.org/10.1007/BF02294241

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika, 61*, 165-170. https://doi.org/10.2307/2334299

Zhang, Y., Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a computerized adaptive testing for internet addiction. *Frontiers in Psychology*, *10*, 1-12. https://doi.org/10.3389/fpsyg.2019.01010

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ω_H: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

APÉNDICES

APÉNDICE 1: Código de simulación en R

```
############################## Este código permite reproducir el experimento de
simulaciones Monte Carlo.
# La instalación de R incluye ciertas funciones del programa base, pero se utilizarán
ciertas funciones adicionales de los paquetes tidyverse y TAM.
# Si no se cuenta con alguna de ellas, es posible instalarlas con la función
install.packages("tidyverse") y install.packages("TAM").

############################## Aquí se designan los paquetes a utilizar
library(tidyverse) #funciones para la manipulación de datos
library(TAM) #funciones para la estimación de parámetros en el modelo Rasch

############################## El experimento Monte Carlo implica la generación de datos
aleatorios; por ello, su replicabilidad require del establecimiento de una semilla
############################## La semilla permite fijar el dispositivo generador de
números aleatorios para que siempre otorgue el mismo resultado
set.seed(1234567)

############################## En las siguientes líneas de código se crea una función
(obs.resp) que permite generar respuestas a ítems dicotómicos sobre la base del modelo
Rasch.
obs.resp <- function(items,personas){
  mI <- t(replicate(length(personas),items))
  mP <- replicate(length(items),personas)
  mPI <- mP-mI
  prob.resp <- exp(mPI) / (1+(exp(mPI)))
  obs.resp <- matrix( sapply( c(prob.resp), rbinom, n = 1, size = 1),
                      ncol = length(items))}

############################## En esta sección se crean las listas necesarias para
almacenar los resultados de cada iteración.
confiabilidadg <- list()
confiabilidadp <- list()
final<- list()

############################## Aquí se crea un objeto llamado reps que permite delimitar
con facilidad las réplicas que se deseen realizar.
reps <- 1000 #réplicas

############################## Aquí se crea un objeto llamado distancia que contiene los
cuatro niveles de la variable independiente: distancia entre medias.
distancia <- c(0.5,1.0,1.5,2.0)

############################## Esta función almacena el tiempo de inicio del experimento.
start.time <- Sys.time()

############################## Para realizar el experimento se requieren varias
iteraciones de un mismo cálculo, ello es posible al delimitar un loop.
############################## En esta primera parte del loop, se establece que todo
procedimiento dentro de los corchetes debe ser reproducido considerando el número de
objetos que tiene la variable distancia (4)
############################## Esto se realiza para que todos los procedimientos se
apliquen a cada uno de los cuatro niveles de la variable independiente: distancia de
medias.
for (k in 1:length(distancia)) {

  ############################## En esta segunda parte del loop se establece que todo
procedimiento entre corchetes debe realizarse un número de veces similar al establecido
en la variable reps (el número de réplicas previamente designado)
  for (i in 1:reps) { #inicio del loop (aquí se pueden delimitar el número de réplicas)

    ############################## En esta sección se delimitan los parámetros de
habilidad de personas para el primer grupo considerando una distribución normal
(M=0,DE=1).
    grupo1 <- rnorm(2500,0,1)

    ############################## Aquí se realiza el mismo procedimiento para el Segundo
grupo considerando una distribución normal (M=k,DE=1, en donde k es cada valor de la
variable distancia)
    grupo2 <- rnorm(2500,distancia[k],1)
```

```
    ############################## En esta sección se delimitan los parámetros de
dificultad de ítems.
    ############################## Primero se generan los parámetros para los ítems en
común (16 en total).
    icomun <- seq(-0.5,2.5,length.out=16)

    ############################## Luego se generan los parámetros para los ítems únicos
del grupo 1.
    items1 <- seq(-2.5,-0.5,length.out=25)

    ############################## Ahora se juntan los parámetros de los ítems en común y
los únicos del grupo 1.
    items1 <- sort(unique(c(items1,icomun)))

    ############################## Aquí se generan los parámetros para los ítems únicos
del grupo 2.
    items2a <- seq(2.5,(2.5+distancia[k]),length.out=(distancia[k]*12+1))
    items2b <- seq((-2.5+distancia[k]),-0.5,length.out=(24-distancia[k]*12+1))

    ############################## Ahora se juntan los parámetros de los ítems en común y
los únicos del grupo y se remueven los insumos previos.
    items2 <- sort(unique(c(items2a,items2b,icomun)))
    remove(items2a,items2b)

    ############################## Con ambos parámetros (habilidad y dificultad) ahora es
posible utilizar la función previamente desginada para generar bases de datos de
respuestas a ítems dicotómicos.
    ############################## Esto se realiza tanto para el grupo 1 como para el
grupo 2
    data1 <- as.data.frame(obs.resp(items1,grupo1))
    colnames(data1) <- items1
    rownames(data1) <- seq(1,2500)
    data2 <- as.data.frame(obs.resp(items2,grupo2))
    colnames(data2) <- items2
    rownames(data2) <- seq(2501,5000)

    ############################## Ahora, ambas bases de datos se integran en una sola
que contiene las respuestas de ambos grupos.
    integrada<- bind_rows(data1, data2)
    integrada<-integrada[ , order(names(integrada))]

    ############################## Una vez se tiene la base de datos, es posible realizar
el cálculo de parámetros
    ############################## La función tam.jml() permite calcular parámetros de
dificultad ítems y en su última iteración estima el WLE para los parámetros de habilidad
de personas.
    MMLEg <- tam.mml(integrada)

    ############################## Entre todos los resultados que otorga esta función,
uno de ellos es el cálculo del índice de confiabilidad de la separación de personas.
    ############################## El resultado es almacenado en una de las listas que se
crearon al inicio del código.
    confiabilidadg[[i]] <- MMLEg$EAP.rel

    ############################## El siguiente paso es calcular la confiabilidad
obtenida a partir de un test on target considerando a toda la población.
    ############################## Para ello, primero se integran los parámetros de
habilidad previamente establecidos.
    pob <- sort(c(grupo1,grupo2))

    ############################## Ahora se calcula el promedio de toda la población con
el fin de realizar un test on target alrededor de este promedio.
    items <- mean(pob)

    ############################## En esta sección se simulan los parámetros de
dificultad de los ítems que componen el test.
    items<- seq((-2.5+items),(2.5+items), length.out=40)

    ############################## Ahora se genera una base de datos con respuestas a
ítems dicotómicos, utilizando la función obs.resp previamente definida.
    data <- as.data.frame(obs.resp(items,pob))

    ############################## Después se realiza la estimación de parámetros de
habilidad y dificultad (similar a como se realizó en los tests on target para grupos)
    MMLEp <- tam.mml(data)
```

```
    ############################## La confiabilidad es almacenada en otra lista.
    confiabilidadp[[i]] <- MMLEp$EAP.rel
  } #este corchete marca el final del loop de las réplicas, es decir, ya se terminaron
las réplicas designadas, en este caso, 1000.

  ############################## Los resultados de las 1000 réplicas son almacenados en
una sola lista denominada final.
  final[[k]] <- as.data.frame(cbind(confiabilidadg,confiabilidadp))
  final[[k]]$dist_m <- rep(distancia[k], reps)
  colnames(final[[k]])<- c("confiabilidad_g","confiabilidad_p","distancia")
} #este corchete marca el final del loop de las categorías de la variable independiente:
distancia de medias, es decir, para cada uno de los cuatro niveles se realizarán 1000
réplicas.


############################## Una vez obtenidas todas las réplicas para cada nivel de la
variable distancia de medias, los resultados se integran en una sola base de datos final
final <- bind_rows(final)

############################## A cada observación se le genera un identificador y se
ordenan las columnas
final <- final %>% mutate(id = 1:nrow(final)) %>%
select(id,distancia,confiabilidad_g,confiabilidad_p) %>% sapply(unlist)

############################## Esta función almacena el tiempo de culminación del
experimento.
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken<- paste0("Inicio: ", start.time," Final: ",end.time," Tiempo empleado:
",time.taken)

############################## Finalmente, estas líneas de código exportan los resultados
en formato csv; y un archivo txt que indica los tiempos de la simulación.
write.csv(final,file="Resultado de simulación.csv",row.names=FALSE)
write(time.taken, file = "Tiempo empleado para la simulación.txt")
```

APÉNDICE 2: Código de análisis en R

```
############################## Este código permite reproducir los análisis del
experimento Monte Carlo.
#Se recomienda utilizar la extensión de R: RStudio
#Última versión de R en la que se ejecutó este script: 3.5.1 (2018-07-02)

############################## Aquí se designan los paquetes a utilizar
library(tidyverse) #funciones para la manipulación de datos
library(reshape2) #más funciones para la manipulación de datos
library(psych) #funciones para las ciencias del comportamiento: permite ejecutar la
función describe()
library(ggpubr) #funciones para crear gráficos de ggplot2 con formato de publicación
científica
library(car) #funciones relacionadas al análisis de regresión (Levene test)
library(PMCMRplus) #funciones relacionadas a las comparaciones múltiples (Hartley test)
library(WRS2) #Anova robusto y pruebas post hoc robustas

############################## En esta sección se carga la base de datos (output del
experimento de simulaciones)
data<- read.csv("Resultado de simulación.csv",stringsAsFactors = FALSE)

########################################################## Análisis exploratorio:
############################################## Estadísiticos descriptivos
############### Modificaciones previas a los datos
data <- split(data[,c(3,4)],data$distancia)

############### Media, desviación estándar, media recortada, valor mínimo, valor máximo
descriptivos<- data %>% map(describe) %>% bind_rows %>% arrange(vars)%>%
               mutate(vars=replace(vars, vars==1, "Grupos"),
                      vars=replace(vars, vars==2, "Población"),
                      distancia=rep(c(0.5,1.0,1.5,2.0),2)) %>%
               select(vars,distancia,n,mean,sd,trimmed,min,max) %>%
               rename(Target=1)

############### Intervalos de confianza
ICg <- list()
ICp <- list()
for (i in 1:length(data)) {ICg[[i]]<- quantile(data[[i]]$confiabilidad_g, probs =
c(0.025, 0.975))
                           ICp[[i]]<- quantile(data[[i]]$confiabilidad_p, probs =
c(0.025, 0.975))}
ICg <- as.data.frame(matrix(unlist(ICg), ncol = 2,  byrow = TRUE)) %>% rename(LI=1,LS=2)
ICp <- as.data.frame(matrix(unlist(ICp), ncol = 2,  byrow = TRUE)) %>% rename(LI=1,LS=2)
IC<- rbind(ICg,ICp)
descriptivos<- cbind(descriptivos,IC) %>% select(Target, distancia, n, mean, LI, LS,
everything())
remove(ICg,ICp,IC)

############### Exportar tabla de estadísticos descriptivos
write.csv(descriptivos,"estadísticos descriptvos.csv",row.names = F)

################################################ Gráficos
############### Modificaciones previas a los datos
data <- read.csv("Resultado de simulación.csv",stringsAsFactors = FALSE)
data <- melt(data,id.vars = c("distancia","id"))
data <- data %>%
        mutate(variable = case_when(variable == "confiabilidad_g" ~ "Grupos",
                                    variable == "confiabilidad_p" ~ "Población")) %>%
        rename(target=variable, confiabilidad=value,distancia=distancia) %>%
        mutate(target=as.character(target),distancia=as.character(distancia)) %>%
        mutate(distancia=sprintf("%.1f", round(as.numeric(data$distancia),1)))

############### Boxplot / Gráfico de cajas y bigotes
box<- ggboxplot(data, x = "distancia", y = "confiabilidad",color =
"target")+ylab("Confiabilidad")+xlab("Distancia entre medias aritméticas")
box<- ggpar(box,legend.title = "Target",legend = "right",palette =
c("#00CED1","#778899"),ylim=c(0.85, 0.925),yticks.by = 0.01)

############### Gráfico lineal de comparación entre medias
line <- ggline(data, x = "distancia", y = "confiabilidad", color = "target",
               add = "mean_se")+ylab("Confiabilidad")+xlab("Distancia entre medias
aritméticas")
```

```
line <- ggpar(line,legend.title = "Target",legend = "right",palette =
c("#00CED1","#778899"),ylim=c(0.85, 0.925),yticks.by = 0.01)

############### Exportar gráficos como png
ggsave("box.png", plot=box, width = 9, height = 6)
ggsave("line.png", plot=line, width = 9, height = 6)

########################################################## Análisis inferencial:
############################################## Supuestos de ANOVA factorial
#################### Modificaciones previas a los datos
data1<- split(data,data$target)
#dataal<- split(data, data$distancia)
datag<- split(data1[[1]],data1[[1]]$distancia)
datap<-split(data1[[2]],data1[[2]]$distancia)
datax<- c(datag,datap)

############################################### Normalidad: Shapiro-Wilk test
shapiro.w<- list()
shapiro.pvalue<- list()
for (i in 1:8) {shapiro.w[[i]]<-  shapiro.test(x = datax[[i]]$confiabilidad)$statistic
                shapiro.pvalue[[i]]<-  shapiro.test(x =
datax[[i]]$confiabilidad)$p.value}
shapiro <-
data.frame(cbind(as.numeric(unlist(shapiro.w)),as.numeric(unlist(shapiro.pvalue))))%>%
          rename(W=1,pvalue=2)
remove(shapiro.w,shapiro.pvalue)
shapiro<- cbind(descriptivos %>%select(Target,distancia),shapiro)

############### Exportar pruebas de normalidad
write.csv(shapiro,"pruebas de normalidad.csv")

############################################### Normalidad: QQplot
qq<- list()
for (i in 1:4) {
  qq[[i]] <- ggplot(datax[[i]], aes(sample=confiabilidad),color=target)+
  geom_qq_line(size=0.8,color="#778899")+
  theme_classic()+
  ylab("Confiabilidad")+
  xlab("Distribución teórica")+
  geom_point(shape=21,colour = "#1E90FF", fill="#00CED1",stat = "qq",alpha =
3/10,size=3)+
  scale_y_continuous(breaks =
round(seq(min(datax[[i]]$confiabilidad),max(datax[[i]]$confiabilidad),length.out =
5),3))}
for (i in 5:8) {
  qq[[i]] <- ggplot(datax[[i]], aes(sample=confiabilidad),color=target)+
  geom_qq_line(size=0.8,color="#778899")+
  theme_classic()+
  ylab("Confiabilidad")+
  xlab("Distribución teórica")+
  geom_point(shape=21,colour = "#DCDCDC", fill="#778899",stat = "qq",alpha =
3/10,size=3)+
  scale_y_continuous(breaks =
round(seq(min(datax[[i]]$confiabilidad),max(datax[[i]]$confiabilidad),length.out =
5),3))}

qq<- ggarrange(qq[[1]],qq[[5]],qq[[2]],qq[[6]],qq[[3]],qq[[7]],qq[[4]],qq[[8]],ncol =
2,nrow=4,labels = c("Distancia de 0.5","","Distancia de 1.0","","Distancia de
1.5","","Distancia de 2.0",""),
                  label.x=0,heights = 1.2)
remove(data1,datax,datag,datap)

############### Exportar qqplots
ggsave("qqplots.png", plot=qq, width = 8, height = 12,limitsize = F)

############################################## Homogeneidad de varianzas: Levene
options(scipen=999)
levene<- list()
levene[[1]] <- leveneTest(data$confiabilidad, interaction(data$target,data$distancia),
center = mean)
levene[[2]] <- leveneTest(data$confiabilidad, interaction(data$target,data$distancia),
center = median)
for (i in 1:2) {
  levene[[i]]<- as.numeric(unlist(levene[[i]]))
  levene[[i]]<- levene[[i]][!is.na(levene[[i]])]
```

```
  levene[[i]]<- round(levene[[i]],3)
}
levene <- data.frame(matrix(unlist(levene),ncol=4,byrow = TRUE))%>%
  rename(df1=1,df2=2,F=3,p.value=4) %>%
  mutate(p.value=replace(levene,p.value==0,"<.001*"),
         centrado=c("Media","Mediana"))%>%
  select(centrado,F,df1,df2,p.value)%>%
  mutate(p.value=as.character(p.value))

#################### Homogeneidad de varianzas: Hartley test
hartley <- list()
for (i in 1:8) {hartley[[i]]<- datax[[i]][,4]}
hartley <- hartleyTest(hartley)
hartley <- t(as.data.frame(c(hartley$statistic,hartley$parameter[1],hartley$p.value)))
rownames(hartley)<-NULL
colnames(hartley)<- c("Fmax","df","pvalue")
hartley <- data.frame(hartley)

############### Exportar pruebas de homogeneidad de varianzas
write.csv(levene,"pruebas de homogeneidad de varianzas - levene.csv",row.names = F)
write.csv(hartley,"pruebas de homogeneidad de varianzas - hartley.csv",row.names = F)

############################################### Anova factorial robusto
#################### Modificaciones previas a los datos
datanova<- as.data.frame(data[,c(1,3,4)]) %>%
  mutate(distancia=as.factor(distancia),
         target=as.factor(target))
datanovapost<- split(datanova,datanova$distancia)

#################### ANOVA factorial robusto
AnovaFr<- t2way(confiabilidad ~ target*distancia, data = datanova, tr = 0.1)
AnovaFr <- data.frame(Fvalue=c(AnovaFr$Qa,AnovaFr$Qb,AnovaFr$Qab),
           pvalue=c(AnovaFr$A.p.value,AnovaFr$B.p.value,AnovaFr$AB.p.value))%>%
           mutate(variable=c("Target","Distancia","Interacción"))%>%
           select(variable,Fvalue,pvalue)

#################### Post hoc
AnovaFrpost<- list()
for (i in 1:4) {AnovaFrpost[[i]]<-  lincon(confiabilidad ~ target, data =
datanovapost[[i]], tr = 0.1)
                AnovaFrpost[[i]]<-AnovaFrpost[[i]]$comp}
AnovaFrpost<- data.frame(cbind(distancia=c(0.5,1.0,1.5,2.0),
                               t(matrix(descriptivos$trimmed,ncol = 4,byrow=T)),
                               data.frame(matrix(unlist(AnovaFrpost),ncol=6,byrow = T))
%>%
                               select(3:6))) %>%

rename(distancia=1,media_recortada_g=2,media_recortada_p=3,psi=4,LI=5,LS=6,pvalue=7)%>%
              mutate(pvalue=replace(AnovaFrpost,pvalue==0,"<.001*"),
                     pvalue=as.character(pvalue))
remove(datanova,datanovapost)

############### Exportar resultados de ANOVA y post hoc
write.csv(AnovaFr,"Anova factorial robusto.csv",row.names = F)
write.csv(AnovaFrpost,"Pruebas post hoc.csv",row.names = F)

#################### ANOVA factorial convencional
#Aquí se presenta el análisis ANOVA factorial convencional solo para indicar que los
resultados son similares al ANOVA robusto
#Este análisis no es parte del experimento; por ello, su output no es exportado
#Para realizar el análisis, borrar el símbolo "#" antes de las siguientes líneas de
código:
#anova <- aov(confiabilidad ~ target + distancia + target*distancia, data = datanova)
#summary(anova)

############################################### Tamaño del efecto
efecto <- data.frame(cbind(distancia=c(0.5,1.0,1.5,2.0),
                     t(matrix(descriptivos$mean, ncol=4 ,byrow = T)))) %>%
          rename(grupos=2,población=3) %>%
          mutate(diferencia=grupos-población)

############### Exportar resultados de tamaño del efecto
write.csv(efecto,"Tamaño del efecto.csv",row.names = F)
```