

Modelos y algoritmos no paramétricos para Test Adaptativos Informatizados Multidimensionales.

Responsable: Mario Luzardo

Equipo: Leonardo Moreno, Darío Padula y Cecilia Marconi.

a) Resumen (no más de una página)

El proyecto tiene por objetivo estudiar la performance de los Test Adaptativos Informatizados (TAI) cuando los ítems se han calibrado mediante modelos no paramétricos multidimensionales de Teoría de Respuesta al Ítem (TRI). Se trabajará con el modelo de Ramsay en su forma multidimensional y el modelo isótono multidimensional desarrollado por Luzardo (2017).

Para ambos modelos se implementarán algoritmos de selección de ítems en tests adaptativos informatizados tanto para ítems dicotómicos como politómicos multidimensionales.

Será necesario deducir la estimación de la información de Fisher de forma no paramétrica para ambos modelos multidimensionales en un punto dado del rasgo. También se encontrará un estimador no paramétrico de la información de Kullback –Leibler.

Para estudiar la adecuación de los estimadores y algoritmos encontrados se compararán los test adaptativos informatizados multidimensionales no paramétricos mediante los dos modelos, estudiando la exactitud de la estimación, el sesgo, la cantidad de ítems necesarios para un error predeterminado y la performance de la estimación ante diversos procedimientos de arranque del test. Se estudiará también el control de exposición de los ítems para analizar la conservación de la calidad del banco de ítems.

En una segunda etapa se compararán los test adaptativos informatizados multidimensionales usando los procedimientos no paramétricos hallados versus los basados en modelos paramétricos tanto en sesgo, error de estimación, control de exposición y métodos de arranque.

b) Presentación y antecedentes de la temática de investigación y fundamentación de su importancia.

Desde hace décadas los TAI han gozado de un gran avance tanto teórico como práctico, ya que tienen varias ventajas ante los test tradicionales y en los últimos 25 años ha crecido su importancia en la evaluación educativa de gran escala.

Específicamente, un test adaptativo informatizado (TAI) tiene como principal característica que sus ítems se presentan y responden mediante un ordenador y que se adapta al nivel de competencia progresivo que va manifestando la persona. Entre las ventajas se incluyen la mayor precisión en la estimación del rasgo, el menor tiempo para su aplicación y su flexibilidad.

Dentro de los TAI sigue siendo tradicional el uso de test unidimensionales, aunque hay, desde ya hace bastante tiempo, TAI multidimensionales basados en modelos paramétricos.

Un aspecto teórico muy importante reside en el hecho que en los TAI se viola el principio de independencia local, que es basamento de la estimación del rasgo en la TRI. Esto se debe a que la selección de los ítems en un paso depende de la respuesta a los ítems anteriores. Esto podría comprometer la consistencia de los estimadores. Sin embargo, Chang, H.-H., & Ying, Z. (2009) demuestran la consistencia y la normalidad asintótica del estimador $\hat{\theta}_n$ por máxima verosimilitud en el caso de TAI en términos de martingalas.

De los componentes de un TAI el más importante, sin dudas, es el algoritmo de selección de ítems, que se usará para elegir qué ítem del banco se utilizará en cada paso para cada examinado durante el test. Dada la confluencia entre los avances psicométricos realizados desde la TRI y los avances técnicos en el campo de la informática, se han desarrollado instrumentos informatizados para presentar únicamente los ítems que resultan altamente informativos para estimar el nivel de cada sujeto en un determinado rasgo (Olea y Ponsoda, 1996). El método más utilizado es el criterio de máxima información de Fisher propuesto por Lord (1980).

Debemos comentar que el método de máxima información de Fisher tiene dos importantes limitaciones: la primera es que los ítems con alta discriminación son elegidos primero, y a menudo, los ítems de baja discriminación, aunque tengan una dificultad adecuada, a veces no son elegidos nunca. La segunda desventaja es que al inicio del test la habilidad es mal estimada.

Ante esto Chang & Ying, (1999) proponen el método a-estratificado el cual usa ítems con baja discriminación al inicio del test cuando la estimación es poco precisa y reservan los ítems de alta discriminación para las etapas posteriores. Con esto logran tener menos tasas de exposición de los ítems lo que preserva el banco. Más adelante, Chang, Qian, & Ying (2001) desarrollaron el método a-estratificado con b-bloques para considerar el caso donde los parámetros de discriminación y de dificultad están correlacionados. Chang & Ying, (2008) demuestran que los ítems con alta discriminación causan saltos grandes en los estimadores de la habilidad cuando hay pocos ítems administrados.

Otro método tradicional para los algoritmos de selección de ítems se basa en maximizar el índice de Kullback-Leibler (Chang & Ying, 1996). Chang & Ying (1996 y 2008) demostraron que seleccionar los ítems mediante el índice de Kullback-Leibler al inicio del test permite estimar con mejor exactitud el rasgo que el método de Fisher.

Los primeros test basados en datos multidimensionales pueden encontrarse en Bloxom y Vale (1987), y Tam (1992) que usan un algoritmo bayesiano elaborado por Owen (1975).

Desde el punto de vista metodológico, se ha investigado bastante el efecto de la multidimensionalidad, por ejemplo, en un principio se estudió el efecto de suponer la unidimensionalidad cuando el rasgo realmente es multidimensional (Ackerman, 1989,1991; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Folk & Green, 1989; Harrison, 1986; Reckase, 1979; Reckase, Ackerman, & Carlson, 1988; Way, Ansley, & Forsyth, 1988; Yen, 1984). Sin embargo, faltan desarrollos metodológicos que impliquen modelos no paramétricos multidimensionales.

Los TAI multidimensionales tienen dos ventajas claras ante los unidimensionales, la primera es garantizar la cobertura de las distintas áreas de contenidos que generalmente en los test unidimensionales se soluciona obligando a que participe en el test cierta cantidad fija de ítems de cada área. Este enfoque tiene inconvenientes cuando alguna de las áreas de contenido está relacionada con la dificultad de los ítems, incidiendo en la eficiencia de las mediciones.

Se puede enfocar este problema desde otras ópticas, por ejemplo, Chang & van der Linden (2003) y van der Linden & Chang (2003) proponen usar programación 0–1 con el método a-estratificado para así balancear los contenidos y mejorar la exactitud. Yi & Chang (2003) y Leung, Chang, & Hau (2003) proponen incorporar la habilidad al método anterior. Este tipo de enfoque tiene problemas computacionales importantes.

En cambio, cuando se utilizan TAI multidimensionales se incorpora información simultánea de todas las dimensiones. Esto ayuda a lograr un adecuado cubrimiento de las áreas de contenido para cada examinado.

La segunda ventaja de los TAI multidimensionales es que son mucho más eficientes en la estimación del rasgo que los TAI unidimensionales, en particular cuando los rasgos medidos están correlacionados, esto logra obtener mayor precisión y hacer los test más cortos.

Respecto a la selección de ítems, Segall (1996, 2001) propone un procedimiento de selección para TAI multidimensionales que incorpora la distribución a priori de la distribución conjunta de la habilidad. Presenta estimadores de máxima verosimilitud y modal Bayesiano para rasgos de dimensión $p > 1$. Este método propone maximizar el determinante de la información de Fisher y es el que produce la menor varianza para el estimador de la habilidad. Este método fue llamado por Mulder & van der Linden, (2009) como método D-óptimo. Como con este método siguen resultando vigentes las limitaciones del método de Fisher ya mencionadas para el caso unidimensional, Veldkamp & van der Linden (2002) proponen usar la integral múltiple ponderada de la información de Kullback-Leibler. Este método es computacionalmente exigente y difícil de implementar. Wang, Chang, & Boughton (2011a, 2011b) en tanto, obtienen resultados sobre la relación del índice de Kullback-Leibler y los parámetros de los ítems.

Wang and Chang (2011b) proponen un nuevo método basado en la información de Kullback-Leibler con actualización Bayesiana, que luego Wang, Chang, and Boughton (2013b) generalizan.

Últimamente se ha trabajado en combinar los test habituales con test de diagnóstico, logrando la multidimensionalidad para permitir obtener perfiles más informativos (Mulder & van der Linden, 2009).

Con respecto al otro aspecto que hemos mencionado que es el control de restricciones, Cheng, Chang & Yi (2007) se enfocan en el balanceo de contenidos. Cheng and Chang (2009) proponen para tratar con múltiples restricciones el índice de máxima prioridad. Más tarde, Cheng, Chang, Douglas, & Guo (2009) lo combinan con un diseño estratificado para el control de la exposición. Recientemente, Zheng, Chang, & Chang, (2013) desarrollaron un método para el balance de contenidos en TAI multidimensionales.

Ya entrando en los modelos, aspecto que es central a este proyecto de investigación, para los TAI es clave el modelo de TRI con que se calibraron los ítems del banco. Este aspecto influye en todos los mencionados ya que la estimación de la habilidad, los métodos de selección de ítems, arranque y balance de contenidos dependen del modelo de TRI utilizado.

El enfoque más popular de la teoría TRI en el caso unidimensional es usar modelos paramétricos, como los modelos logísticos de uno, dos o tres parámetros, siendo el modelo de Rasch actualmente el más popular. La literatura sobre estos modelos es extensa, particularmente Lord (1980); Hambleton, Swaminathan y Rogers (1991); Fisher y Molenaar (1995); Van der Linden y Hambleton (1997); Boomsma, Van Duijn y Snijders (2001); y Baker y Kim (2004). Estos modelos determinan la forma de la curva característica del ítem (CCI), dependiendo de un número fijo y pequeño de parámetros. Sin embargo, estos modelos no toman en consideración la no monotonía de los ítems y formas funcionales no logísticas, además son poco flexibles (Douglas, 1997; Douglas y Cohen, 2001; Ramsay, 1991).

En los modelos paramétricos, los métodos para estimar la CCI incluyen la estimación por máxima verosimilitud conjunta, la verosimilitud marginal y la estimación condicional de máxima verosimilitud, así como la estimación bayesiana; pero si se violan los supuestos de unidimensionalidad e independencia local, las estimaciones de los parámetros de los ítems y la habilidad son pobres. Por tanto, los modelos paramétricos tienen limitaciones importantes, por ejemplo, no toman en cuenta elementos tales como la no monotonía de los ítems, así como otros apartamientos de la forma propuesta. Además, se han observado que las estimaciones del modelo de tres parámetros son inestables y que en todos los modelos los parámetros tienen covarianzas muestrales grandes (Lord, 1980).

Para lograr una mayor flexibilidad se han desarrollado varios métodos alternativos, basados en enfoques cuasi-paramétricos o no paramétricos. Se han realizado importantes investigaciones sobre aquellos modelos que no se basan únicamente en un enfoque paramétrico: como el modelo de spline parcial, que es en parte paramétrico y en parte no paramétrico (Wahba, 1990) o modelos basados en una combinación lineal de funciones básicas. Uno de estos métodos, la teoría de la respuesta de ítems no paramétrica, demostró ser más flexible. Los primeros modelos no paramétricos fueron propuestos por Mokken (1971), Niemoller y Van Schuur (1983), Mokken y Lewis (1982), Sijtsma (1988) y Giampaglia (1990). Más recientemente, Mokken (1997), Sijtsma (1998, 2001), Molenaar y Sijtsma (2000), Junker (2001) y Junker y Sijtsma (2001) presentaron nuevos resultados.

Desde el enfoque no paramétrico clásico se destacan dos modelos de TRI, que se encuentran en Mokken (1971): el modelo homogéneo monótono y el modelo homogéneo de doble monotonía. El tratamiento matemático para el primero se puede encontrar en Holland y Rosenbaum (1986), Holland (1990), Stout (1987, 1990), Junker (1993) y Ellis y Junker (1997).

Mokken (1971) critica la utilización de modelos paramétricos afirmando que para hacer uso de ellos se necesita un amplio conocimiento de los ítems, y que este conocimiento sea asequible, y existen áreas en que, por razones prácticas, nunca se lograría esto. Para superar esta limitación se busca una alternativa basada solo en restricciones de orden (Luzardo, 2017). En su trabajo Mokken (1971) demostró que si se asume el modelo monótono homogéneo y se cumplen los requisitos previos de unidimensionalidad, independencia local y monotonía, la covarianza entre todos los pares de ítems es no negativa. Como parte de las ventajas de este modelo es que muestra estimadores robustos de la habilidad de las personas y de la dificultad de los ítems y tiene muy buena capacidad predictiva sin outliers.

Además, Rosenbaum (1984) y Holland y Rosenbaum (1986) demostraron la asociación condicional en modelos monótonos de rasgos latentes. Meijer, Sijtsma y Smid (1990)

proporcionaron una comparación cuidadosa entre los modelos paramétricos y no paramétricos, al igual que De Koning, Sijstma y Hamers (2002).

Ramsay (1991) presenta un método novedoso basado en regresión no paramétrica mediante núcleos para obtener una estimación de la CCI para ítems unidimensionales, que tiene como gran aporte introducir un método de análisis de ítems desde la perspectiva funcional. Este enfoque proporciona métodos para estimar las CCI que sean suaves (diferenciables) y al mismo tiempo cumplan ciertos criterios sobre la forma. Este modelo se implementó en el paquete TestGraph (Ramsay, 2000). Este método tiene varias ventajas: la primera de ellas es la no imposición de ninguna restricción sobre la forma funcional de las CCI, y además, el método de estimación es no iterativo, rápido, eficiente y muy fácil de programar. Además, estas técnicas permiten tener herramientas de la TRI para situaciones de muestras de tamaño moderado (Luzardo, 2017).

Douglas (1997) probó la consistencia conjunta para el caso unidimensional, mientras Luzardo y Forteza (2014) presentan una extensión del modelo al caso multidimensional y Luzardo (2017) prueba la consistencia conjunta para el modelo multidimensional bajo condiciones generales. Además, Luzardo (2017) presenta un modelo no paramétrico isótono tanto para el caso de rasgo unidimensional como multidimensional donde la característica del modelo consiste en estimar la inversa de la CCI.

Por tanto, existen los desarrollos teóricos necesarios para construir sobre estos un aporte metodológico en el desarrollo de TAI no paramétricos dado que no hay avances en la materia. En TAI no paramétricos unidimensionales hay básicamente dos referencias: Xu & Douglas (2006) que proponen para ítems que siguen el modelo de Ramsay, un algoritmo de selección de ítems basado en la entropía de Shannon y en la información de Kullback-Liebler, y Luzardo (2019) que propone un algoritmo de selección de ítems basado en la máxima información obtenida a partir del modelo no paramétrico isótono obteniendo mejores resultados ante los métodos anteriores.

c) Descripción del problema de investigación.

Pese al avance y aplicaciones que hay actualmente en torno a los TAI, la inmensa mayoría de ellos se desarrollan bajo el paradigma de los modelos paramétricos y con estimación de la información de Fisher o de la distancia de Kullback-Leibler también de manera paramétrica. No hay aplicaciones de gran escala usando modelos no paramétricos y existen pocos desarrollos teóricos a nivel no paramétrico (Douglas 2006, Luzardo 2019) siendo estos para el caso unidimensional.

A nivel de TAI multidimensionales no paramétricos no hay desarrollos de algoritmos de selección de ítems, procedimientos de arranque, control de exposición y balance de contenido. Tampoco hay comparaciones de su performance con respecto a TAI desarrollados con modelos paramétricos.

Nuestro problema de investigación se relaciona con la posibilidad de aplicar TAI no paramétricos en el caso de ítems multivariados que pueden depender de rasgos correlacionados o no.

Este campo está completamente inexplorado, por lo tanto se deben desarrollar cuatro aspectos:

- a) algoritmos de selección de ítems,
- b) procedimientos de arranque del TAI multidimensional no paramétrico,
- c) métodos para el control de la exposición de los ítems,
- d) métodos de balanceo de contenido.

Cada uno de los aspectos mencionados será abordado para el modelo multidimensional de Ramsay (Luzardo 2017) y para el modelo isótono no paramétrico multidimensional (Luzardo 2019).

A los efectos de poder resolver los puntos mencionados deberán proponerse estimadores no paramétricos (para rasgos multidimensionales) de la información de Fisher y del índice de Kullback-Lieber.

Finalmente, deberá compararse la performance de los TAI multidimensionales no paramétricos versus los paramétricos, basándose en:

- a) error cuadrático medio para el test total
- b) sesgo
- c) longitud del test necesaria para un error prefijado

- d) índice de exposición de los ítems
- e) error cuadrático medio en función de la longitud del test

d) Objetivos generales y específicos del proyecto.

El objetivo general refiere a la generación de conocimiento en TAI multidimensionales que utilizan modelos no paramétricos de TRI en términos de algoritmia, exactitud, longitud y exposición de los ítems; buscando generar además comparaciones en estas medidas con los TAI que utilizan modelos paramétricos de TRI.

Para ello, es necesario abordar los siguientes objetivos específicos:

- a) Desarrollar, mediante diferentes técnicas, algoritmos de selección de ítems en el TAI cuando los ítems siguen modelos no paramétricos multidimensionales.
- b) Desarrollar métodos de estimación del rasgo latente y comprobar su consistencia estadística cuando se utilizan modelos no paramétricos para el TAI.
- c) Comparar la eficiencia de los TAI no paramétricos con los paramétricos en términos de cantidad de ítems administrados, sesgo de las estimaciones, exposición de los ítems y error cuadrático medio.
- d) Implementar algoritmos para el control de la exposición de los ítems para el caso de TAI no paramétricos.
- e) Divulgar la utilización de los TAI multidimensionales que utilizan modelos no paramétricos mediante publicaciones científicas en revistas reconocidas.
- f) Formar recursos humanos capaces de investigar en TAI no paramétricos.

e) Preguntas que busca responder el proyecto e hipótesis, si corresponde.

Las preguntas que este proyecto intenta responder son:

1. ¿Qué algoritmo de selección de ítems en términos de exactitud en la estimación estimada mediante el error cuadrático medio, e igual cantidad de ítems presentados, es el mejor para TAI no paramétricos multidimensionales?
2. Si los ítems siguen un modelo paramétrico multidimensional, ¿cómo se comportan los TAI multidimensionales no paramétricos respecto a los que usan los modelos paramétricos?
3. Si los ítems no siguen un modelo paramétrico multidimensional, ¿qué ganancia en términos de exactitud y sesgo se obtiene al usar TAI no paramétricos multidimensionales ante los paramétricos?
4. ¿Qué diferencias hay en la longitud del test para un error prefijado, al usar TAI no paramétricos versus los paramétricos ante violaciones de los ítems del modelo multidimensional paramétrico?
5. ¿Cómo impacta en la exposición de los ítems y por ende en la calidad del banco el uso de TAI no paramétricos multidimensionales versus los paramétricos ante diferentes modelos de los ítems?

f) Estrategia y metodología de investigación, así como actividades específicas.

La estrategia de investigación se desarrollará en dos componentes. El primer componente se enfoca en la búsqueda y desarrollo de las estrategias para TAI no paramétricos multidimensionales, implica el desarrollo de algoritmos de selección de ítems, criterios de arranque, criterios de finalización del TAI y métodos para conservar la calidad del banco. El segundo componente consistirá en un estudio mediante simulaciones de la performance de los algoritmos hallados y comparación de los TAI no paramétricos multidimensionales versus los TAI multidimensionales paramétricos.

En el segundo componente se usará para las simulaciones en el software R y se utilizará un diseño que considerará 24 poblaciones que serán normales de dimensión 2 y 3 con cuatro diferentes grados de correlación entre las diferentes dimensiones ($\rho = 0, \rho = 0.3, \rho = 0.6, \rho = 0.9$). Observemos que en el caso de dos dimensiones tenemos 4 poblaciones y en el caso de tres dimensiones tenemos 20 poblaciones de diferentes configuraciones de la matriz de correlaciones. Las poblaciones tendrán un tamaño de 5000 sujetos.

Para cada población se simularán las respuestas de los sujetos cuando los ítems siguen tres posibles modelos multidimensionales, dos de ellos (MIRT y logístico) que podrán ser estimados correctamente mediante los modelos paramétricos y no paramétricos y un modelo (logit cúbico multidimensional) que solo puede ser estimado correctamente mediante un modelo no paramétrico, lo que pondrá en evidencia la flexibilidad de estos modelos.

Se obtendrán, por lo tanto, 96 conjuntos de respuestas las cuales serán utilizadas para calibrar los cuatro modelos que consideraremos en el banco de ítems para los TAI multidimensionales. Los modelos para modelizar el banco de ítems serán MIRT, logístico multidimensional, Ramsay multidimensional e isótono multidimensional. A partir de esto, se obtendrán 384 bancos de ítems. Finalmente, para cada uno de los bancos se armará el TAI considerando cada uno de los modelos mencionados. Cada TAI se evaluará a longitud fija 5, 10, 15, 20, 30, 40 y 50 ítems y también a la longitud obtenida para un error prefijado.

Se calculará para cada TAI a los efectos de comparar y obtener conclusiones las siguientes medidas:

- 1) error cuadrático medio de las estimaciones al finalizar el test,
- 2) error cuadrático medio de las estimaciones en cada subtest,
- 3) sesgo de la estimación al finalizar el test,
- 4) sesgo de la estimación en cada subtest.
- 5) longitud del test necesaria para obtener un error prefijado
- 6) índice de exposición de los ítems

Las actividades específicas están referidas en los objetivos específicos del proyecto, ellas son:

- a) Desarrollo de algoritmo de selección de ítems para ítems dicotómicos no paramétricos multidimensionales usando el índice de Kullback–Leibler y criterio de máxima información de Fisher (en el caso multidimensional).
- b) Desarrollo de algoritmo de selección de ítems para ítems dicotómicos multidimensionales con el modelo isótono no paramétrico usando información de Fisher y el índice de Kullback–Leibler.
- c) Desarrollo de criterios de arranque para el TAI multidimensional no paramétrico.
- d) Simular 5000 rasgos de examinados con distribución normal mutivariada (se simularán los casos bidimensionales y tridimensionales) con diferentes grados de correlación. Se simularán 4 cuatro condiciones de correlacion entre pares de rasgos ($\rho = 0, \rho = 0.3, \rho = 0.6, \rho = 0.9$). Con esto obtendremos un diseño con 24 poblaciones simuladas.
- e) Simular las respuestas de los sujetos de cada población según los siguientes modelos dicotómicos:
 - a. Modelo multidimensional normal, MIRT
 - b. Modelo logístico multidimensional
 - c. Modelo logístico cúbico. Este modelo no ajustará a los modelos paramétricos.

Con los puntos c) y d) se obtendrán 96 conjuntos de datos.

- f) Calibrar los modelos paramétricos (MIRT y logístico multidimensional) y los no paramétricos (Ramsay multidimensional e Isótono multidimensional) con los conjuntos de datos hallados en d). Se obtendrán cuatro bancos de ítems para cada conjunto de datos, en total 384 bancos.
- g) Simular el TAI multidimensional para los cuatro bancos de ítems armados con cada conjunto de datos.
- h) Estimar los rasgos con subtest de longitud 5, 10, 15, 20, 30, 40 y 50 ítems.
- i) Estimar los rasgos con los criterios de parada prefijados.

- j) Estimar la longitud media del TAI para criterios de exactitud predeterminados. También se estimará la densidad de la distribución de las longitudes.
- k) Comparación de los resultados en términos de exactitud, cantidad de ítems utilizados, exposición de los ítems y sesgo para el test completo y para los subtests.
- l) Comparación de los métodos de selección de ítems usando modelos no paramétricos versus modelos paramétricos.
- m) Comparar la performance del TAI no paramétrico según varios procedimientos de arranque.
- n) Explorar procedimientos de selección de ítems y estimación de rasgos alternativos como ser: estimación Bayesiana usando modelos no paramétricos, y selección de ítems usando modelos no paramétricos politómicos.
- o) Mostrar las diferencias entre los modelos paramétricos, el modelo de Ramsay con el modelo isótono no paramétrico.
- p) Estudiar la viabilidad de implementación de los algoritmos desarrollados.
- q) Redacción de informe final.

g) Descripción de los beneficios esperados de los resultados tanto en términos académicos como en términos sociales, económicos, productivos, si corresponde.

Se espera que con los resultados obtenidos se incremente el uso de TAI no paramétricos, tanto multidimensionales como unidimensionales, debido a que tienen ventajas comparativas en términos de flexibilidad y esperamos demostrar que también en términos de exactitud y de simplicidad en la programación de los algoritmos.

Por otra parte, el proyecto prevé la participación de estudiantes en formación (maestría y doctorado) donde sus áreas de estudio están comprendidas dentro de las temáticas que aborda la investigación. La estudiante de doctorado tiene su proyecto de tesis en concordancia con el tema de este proyecto, en tanto el estudiante de maestría, que ya tiene una maestría previa, se continúa y que tiene gran experiencia en TRI, se perfeccionará en TAI multidimensionales lo que se alinea con su proyecto de formación en ciencia de datos.

En términos sociales recordemos que a nivel nacional existen antecedentes muy valiosos de aplicación de TAI a escala nacional y con fines de evaluación educativa. Los test adaptativos que se usan en Uruguay tuvieron su origen y desarrollo en el marco de convenios entre la División de Investigación, Evaluación y Estadística (DIEE) de la Administración Nacional de Educación Pública (ANEP), donde se realizó la programación de los algoritmos en el sistema de evaluación en línea de aprendizajes (SEA), y la Facultad de Psicología de la Universidad de la República de Uruguay (Udelar). Específicamente, en Uruguay se destaca la aplicación de metodologías TRI en la evaluaciones adaptativa de inglés en educación Primaria y educación Media; la aplicación reciente de TAI, por parte de la DIEE, en matemática, ciencia y lengua; y en pruebas diagnósticas de competencias al ingreso de la Universidad (Rodríguez Morales, 2017). La evaluación adaptativa de inglés en el sistema educativo uruguayo se realizó por primera vez a fines de 2014, donde participaron estudiantes de inglés de 4to, 5to y 6to año de escuelas públicas de Uruguay que estudian inglés tanto a través del Programa del Departamento de Segundas Lenguas y Lenguas Extranjeras de la ANEP (programa presencial), como del Programa Ceibal en Inglés de Plan Ceibal (programa por videoconferencia). Fue diseñada para evaluar los aprendizajes de todos los alumnos de 4º a 6º de Educación Primaria y de los alumnos de 1º a 3º de Educación Media Básica. Esta evaluación se ha implementado anualmente y en 2018, en su quinta edición, algo más de 74.000 alumnos de Primaria y Media aplicaron el TAI. Otro antecedente en el contexto nacional es la experiencia de aplicación de TAI en la Udelar, específicamente, en el Centro Universitario Regional Este (CURE). En este caso para la aplicación de TAI en matemática con el objetivo de tener una evaluación diagnóstica de las nuevas generaciones de estudiantes que ingresan al CURE (Rodríguez, Pérez y Luzardo, 2017).

Por lo tanto, esperamos que lo hallado en esta investigación se pueda continuar a nivel nacional utilizando TAI paramétricos unidimensionales y que se comience a utilizar TAI

multidimensionales, en particular modelizados mediante TRI no paramétrica, por las ventajas que esto ofrece.

h) Personal asignado al proyecto y personal a contratar; detalle de las tareas a realizar por cada integrante.

Mario Luzardo: (responsable): Estará a cargo de la coordinación y gestión del proyecto. Diseñará los planes de análisis en consulta con los integrantes del equipo para la consecución de los objetivos. Desarrollará los algoritmos necesarios. Coordinará con los demás integrantes del equipo la distribución de las tareas. Se encargará de la redacción de informes y artículos y difusión de resultados.

Leonardo Moreno: Asesorará en el desarrollo de los algoritmos y en los procesos de pauta de simulación. Supervisará los planes de análisis y análisis psicométricos. Colaborará en la redacción de informes, artículos y difusión de resultados.

Darío Padula (estudiante de maestría): Participará en el desarrollo de los algoritmos y en el estudio de simulación, proveerá los códigos en R para las simulaciones. Colaborará en la redacción de informes, artículos y difusión de resultados.

Cecilia Marconi (estudiante de doctorado): Colaborará principalmente con las simulaciones en R. Participará en la redacción de informes, artículos y difusión de resultados.

i) Formación de recursos humanos.

Doctorado: El Proyecto se relaciona con el doctorado de Cecilia Marconi en Facultad de Psicología. El tema de trabajo de Cecilia Marconi son los test adaptativos unidimensionales no paramétricos y al tratarse este proyecto de TAI multidimensionales no paramétricos se adecua perfectamente a su proceso formativo.

Maestría: El proyecto se relaciona con la formación de Darío Padula en su segunda maestría en Ciencias de Datos brindada por el Centro de Posgrados y Actualización Profesional (CPAP), Facultad de Ingeniería, Udelar (la maestría está en trámites de aprobación). La formación radica en extraer conocimiento a partir de los datos y en la aplicación de técnicas estadísticas, de aprendizaje automático, manejo de python, R, matlab, bases de datos, procesamiento de grandes volúmenes de datos, análisis de lenguaje natural, etc. Por lo tanto, lo desarrollado en este proyecto se relaciona estrechamente con los objetivos de su maestría.

j) Equipos y materiales, si corresponde. Descripción de equipos y materiales disponibles para el desarrollo del proyecto; en caso de solicitar nuevos equipos y materiales, es necesario fundamentar su necesidad.

El equipo de investigación cuenta con lo necesario para desarrollar el proyecto, tanto a nivel informático como bibliográfico (se cuenta con acceso a las revistas con los artículos que se necesitan).

k) Cronograma de ejecución, especificando los resultados que se espera obtener en cada etapa (siempre que corresponda).

Se presenta el cronograma de ejecución dividido en trimestres. Cada literal en la columna “Actividades” hace referencia a lo mencionado en el punto f sobre la Estrategia y metodología de investigación, así como actividades específicas.

Actividades	Resultado	1	2	3	4	5	6	7	8
-------------	-----------	---	---	---	---	---	---	---	---

a)	Algoritmo	X	X						
b)	Algoritmo	X	X						
c)	Algoritmo		X	X					
d)	Set de datos		X						
e)	Set de datos		X						
f)	Banco de ítems		X	X					
g)	Script de R			X	X	X			
h)	Set de datos					X			
i)	Set de datos					X			
j)	Set de datos					X			
k)	Inferencia estadística						X	X	
l)	Inferencia estadística						X	X	
m)	Inferencia estadística						X	X	
n)	Inferencia estadística						X	X	
o)	Inferencia estadística						X	X	
p)	Documento								X
q)	Documento								X

l) Referencias bibliográficas.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non-compensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, 15, 3-24.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37--48.

Baker, F.B. & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.

Boomsma, A.; Van Duijn, J. & Snijders, A.B. (Eds)(2001). Essays on item response theory. New York: Springer.

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211–222.

Chang, H.-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement*, 25(4), 333–341.

Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73(3), 441–450.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37(3), 1466–1488.

- Chang, H.-H., & van der Linden, W.J. (2003). Optimal stratification of item pools in a -stratified computerized adaptive testing. *Applied Psychological Measurement*, 27(4), 262–274.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3), 213–229.
- Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, 31(6), 467–482.
- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369–383.
- Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted a -stratification for computerized adaptive testing with non-psychometric constraints: balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, 69, 35–49.
- De Koning, E.; Sijstma, K. & Hamers, J.H.M. (2002). Comparison of four IR models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement*, 26 (3), 302-320.
- Douglas, J. (1997). Joint Consistency of Nonparametric Item Characteristic Curve and Ability Estimation. *Psychometrika*, 62, 7-28.
- Douglas, J. & Cohen, A (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25, 234-243.
- Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
- Ellis, J. L. & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, 62, 495-523.
- Fischer, G. H. & Molenaar, I. (Eds.)(1995). *Rasch Models. Foundations, Recent Developments and Applications*. New York: Springer.
- Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-389.
- Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 11, 91-115.
- Hambleton, R.; Swaminathan, H. & Rogers, H. (1991). *Fundamentals of Item Response Theory*, Thousand Oaks: Sage Publications.
- Holland, P.W. (1990). On the sampling theory foundations of item response theory models, *Psychometrika*, 55, 577-601.
- Holland, P.W. & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models, *The Annals of Statistics*, 14, 1523-1543.
- Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models, *The Annals of Statistics*, 21, 1359-1378.
- Junker, B.W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M.A., J. Van Duijn y T. A. B. Snijders (Eds.), *Essays on item response theory* (pp.247-276), New York: Springer.
- Junker, B.W. & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue, *Applied Psychological Measurement*, 25, 211-220.
- Leung, C., Chang, H.-H., & Hau, K. (2003). Computerized adaptive testing: a comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment*, 2(5), 2–15.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: LEA.
- Luzardo, M. & Forteza, D. (2014). *Modelo no paramétrico multidimensional para la estimación de los rasgos y de las curvas características del ítem mediante regresión no paramétrica con núcleos*. Montevideo: CSIC.
- Luzardo, M. & Rodríguez P. (2015). A Nonparametric Estimator of a Monotone Item Characteristic Curve. In: van der Ark L., Bolt D., Wang WC., Douglas J., Chow SM. (eds) *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics*, vol 140. Springer, Cham
- Luzardo, M. (2017). *Contribuciones a la Teoría de Respuesta al Ítem no Paramétrica*. Tesis de Doctorado. Madrid: Facultad de Psicología, Universidad Autónoma de Madrid.
- Luzardo, M. (2019). Item Selection Algorithms in Computerized Adaptive Test Comparison Using Items Modeled with Nonparametric Isotonic Model. In: Wiberg M., Culpepper S., Janssen R., González J., Molenaar D. (eds) *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics*, vol 265. Springer, Cham

- Meijer, R. R., Sijtsma, K. & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, 14, 283-298.
- Mokken, R. J. (1971). A theory and procedure of scale analysis. New York/Berlin: De Gruyter.
- Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6, 417-430.
- Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, 74(2), 273-296.
- Niemoller, K. & W. van Schuur (1983). Stochastic Models for Unidimensional Scaling: Mokken and Rasch. In D. McKay N. Schofield (Eds.), *Data analysis and the social sciences*. London: Frances Printer Ltd.
- Olea, J. & Ponsoda, V. (1996). Test adaptativos informatizados. En J. Muñiz (Ed.), *Psicometría*. Madrid: Universitas, S. A.
- Ramsay, J.O.(1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611-630.
- Ramsay, J.O. (2000). TestGraf: A computer program for nonparametric analysis of testing data. Unpublished manuscript, McGill University.
- Ramsay, J.O. & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, 56, 365-379.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193-203.
- Rodríguez Morales, P. (2017). Creación, desarrollo y resultados de aplicación de pruebas de evaluación basadas en estándares para diagnosticar competencias en Matemática y Lectura al ingreso a la Universidad. *Revista Iberoamericana de Evaluación Educativa*, 10 (1), 89-107.
- Rodríguez, P., Pérez, G. y Luzardo, M. (2017). Desarrollo y aplicación del primer test adaptativo informatizado (TAI) de Matemática para orientar trayectorias en la Universidad. En N. Peré (Comp.) *La Universidad Se Investiga* (pp. 1041-1048). Jornadas de Investigación en Educación Superior. Montevideo: CSE-ANEP.
- Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425-435
- Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331-354.
- Segall, D.O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*, 66(1), 79-97.
- Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores, *Applied Psychological Measurement*, 22, 3-32.
- Sijtsma, K. (2001). Developments in measurement of persons and items by means of item response models. *Behaviormetrika*, 28, 65-94.
- Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: Sage Publications.
- Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.
- Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika*, 55, 293-325.
- van der Linden, W. & Hambleton, R. (1997). *Handbook of Modern Item Response Theory*, New York: Springer-Verlag.
- van der Linden, W.J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, 27(2), 107-120.
- Veldkamp, B.P., & Van Der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67(4), 575-588.
- Wang, C., Chang, H.-H., & Huebner, A. (2011a). Restrictive stochastic item selection methods in cognitive diagnostic CAT. *Journal of Educational Measurement*, 48(3), 255-273.
- Wang, C., Chang, H.-H., & Boughton, K. (2011b). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, 76(1), 13-39.

- Wang, C., Chang, H., & Boughton, K. (2013b). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 37, 99–122.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement*, 12, 239–252.
- Wahba, G. (1990). Spline Models for observational data. Philadelphia: Society for Industrial and Applied Mathematics.
- Xu, X. & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT model, *Psychometrika*, 71, 121–137.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the threeparameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yi, Q., & Chang, H.-H. (2003). α -stratified CAT design with content blocking. *British Journal of Mathematical & Statistical Psychology*, 56, 359–378.
- Zheng, Y., Chang, C.-H., & Chang, H.-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, 22, 491–499.