



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

FACULTAD DE CIENCIAS SOCIALES

ESCUELA DE PSICOLOGIA

**Escalamiento vertical longitudinal en mediciones de Desarrollo Infantil en Chile**

**SEBASTIÁN ANTONIO RODRÍGUEZ ANAIZ**

Profesor Guía: David Torres Iribarra

Tesis presentada a la Escuela de Psicología de la Pontificia Universidad Católica de Chile para  
optar al grado académico de Magíster en Psicología Educacional

**Diciembre, 2020**

**Santiago, Chile**

## Contenido

|                                                                                                     |           |
|-----------------------------------------------------------------------------------------------------|-----------|
| <b>Resumen de la Tesis .....</b>                                                                    | <b>3</b>  |
| <b>Introducción.....</b>                                                                            | <b>3</b>  |
| <b>Antecedentes Teóricos y Empíricos .....</b>                                                      | <b>5</b>  |
| <i>Características de la Medición del Desarrollo Infantil.....</i>                                  | <i>5</i>  |
| <i>Interpretación de resultados sobre desarrollo infantil.....</i>                                  | <i>6</i>  |
| <i>Escala común para el desarrollo infantil – Escalamiento Vertical.....</i>                        | <i>11</i> |
| <i>Mediciones de desarrollo infantil a nivel poblacional y longitudinal: El caso de Chile .....</i> | <i>18</i> |
| <b>Planteamiento del problema.....</b>                                                              | <b>21</b> |
| <b>Objetivo.....</b>                                                                                | <b>22</b> |
| <b>Metodología .....</b>                                                                            | <b>22</b> |
| <i>Datos .....</i>                                                                                  | <i>22</i> |
| <i>Procedimiento Estadístico .....</i>                                                              | <i>23</i> |
| <i>Decisión de la estructura del test.....</i>                                                      | <i>23</i> |
| <i>Decisiones del modelo estadístico.....</i>                                                       | <i>30</i> |
| <i>Análisis psicométricos ítems comunes.....</i>                                                    | <i>30</i> |
| <b>Resultados .....</b>                                                                             | <b>35</b> |
| <i>Comparación de estimaciones ancladas y libres .....</i>                                          | <i>35</i> |
| <i>Comparación de estimaciones entre versiones de anclaje .....</i>                                 | <i>38</i> |
| <b>Conclusión.....</b>                                                                              | <b>41</b> |
| <b>Bibliografía .....</b>                                                                           | <b>44</b> |
| <b>Anexos .....</b>                                                                                 | <b>47</b> |
| <i>Anexo N°1.....</i>                                                                               | <i>47</i> |
| <i>Anexo N°2.....</i>                                                                               | <i>48</i> |
| <i>Anexo N°3.....</i>                                                                               | <i>52</i> |

## **Resumen de la Tesis**

Esta Tesis se focaliza en las interpretaciones longitudinales sobre el desarrollo infantil. Las interpretaciones de puntajes bajo normas no permiten dar cuenta de la característica de trayectoria en el tiempo que posee de forma inmanente el desarrollo infantil. Ante esta problemática se hace necesario generar una escala común del desarrollo infantil por medio de un procedimiento de escalamiento vertical.

Haciendo uso de datos poblacionales y longitudinales chilenos provenientes de la Encuesta Longitudinal de Primera Infancia, se utilizan los resultados del instrumento Inventario de Desarrollo Battelle en sus versiones BDI y BDI-2 Screening para realizar un procedimiento de escalamiento vertical considerando un diseño de ítem común, una estimación concurrente del Modelo de crédito parcial y una estimación de habilidad por medio de EAP.

Los resultados del escalamiento vertical son satisfactorios, permitiendo interpretaciones longitudinales y poblacionales del desarrollo infantil en Chile.

## **Introducción**

El desarrollo humano comprende las interacciones del sujeto con su ambiente que provocan cambios cuantitativos y cualitativos en la forma que tendrá este para relacionarse con su entorno (Vigotski, 1979). De igual forma, a medida que pasa el tiempo, las interacciones que tiene el individuo con su ambiente se basan en las ya adquiridas (Piaget & Inhelder, 2008). Un periodo clave dentro de este ciclo vital es la primera infancia (Bedregal & Pardo, 2004).

Según los postulados de Bedregal & Pardo (2004), se ha reconocido que las niñas y niños que se encuentran en el periodo de la primera infancia (0-8 años) están sujetos a derechos particulares, debido a su condición de seres en desarrollo psicosocial. En este periodo la interacción de los infantes con su ambiente sienta las bases de todo desarrollo posterior. A pesar de que el desarrollo humano es un proceso continuo a lo largo del ciclo vital, es en la primera infancia donde se generan los pilares de la actividad humana posterior.

Sin embargo, a la hora de medir el desarrollo infantil, Raikes (2017) plantea tres aspectos a considerar.

En primer lugar, el desarrollo infantil es holístico, y en su medición se debiesen incluir los cambios psicosociales, físicos y cognitivos (Raikes, 2017). Teorías evolutivas sobre

dominios cognitivos (Flavell, Miller, & Miller, 1993), socioemocionales (Sroufe, 1997) o del lenguaje (Mariscal, 2008) han sido desarrolladas. Además, el desarrollo infantil no sólo implica distintos dominios, sino también interacciones entre estos y patrones de evolución distintos entre cada uno (Raikes, 2017). Por ende, para obtener mediciones precisas sobre el desarrollo infantil, se recomienda utilizar instrumentos multidominios (Fernald, Kariger, Engle, & Raikes, 2009).

En segundo lugar, la cultura y el contexto tienen una profunda influencia en la adquisición y emergencias de las competencias del desarrollo (Raikes, 2017). El desarrollo infantil se ve afectado no sólo por las relaciones próximas que tiene el infante, sino además por los otros sistemas en que se engloban las interacciones directas que tiene la niña o niño (Bronfenbrenner, 1979). A partir de esto, se puede argumentar que las competencias adquiridas por el infante varían según su entorno. La no demostración de una habilidad por parte de la niña o niño se puede deber a la falta de relevancia cultural que se le otorga a dicha habilidad en ese contexto, o a una insuficiencia de estímulos que puede ser interpretada como un rezago en el desarrollo (Raikes, 2017). Debido a esto, los instrumentos de desarrollo infantil no son transculturales, y por ende, las brechas en el desarrollo no implican necesariamente desigualdades en logros en una escala universalmente válida (Bedregal, 2008).

En tercer lugar, el desarrollo infantil es cambiante a lo largo del tiempo. Tal como lo plantea Piaget & Inhelder (2008), durante la vida de un infante habrá distintos estadios, ordenados de forma secuencial. Cada estadio le permitirá realizar distintas tareas, con mayor dificultad a medida que pasa el tiempo. Es así como, a medida que el infante se desarrolla, va asimilando nuevas estructuras para solucionar los obstáculos a los cuales se enfrenta. Es por esto que el desarrollo infantil debe ser entendido como una trayectoria, que crece a medida que aumenta la edad de la niña o niño.

Es en esta complejidad de la medición sobre desarrollo infantil donde se pondrá el foco en el presente trabajo.

Debido al cambio en el tiempo que tiene este constructo, uno de los aspectos a considerar en toda medición es su capacidad de representar el desarrollo infantil. En específico Raikes (2017), remarca que hasta la fecha escasea la investigación de carácter

poblacional y longitudinal que briden evidencia clara de cómo se relacionan los puntajes con los logros que adquieren infantes a lo largo del tiempo.

A su vez, considerando la importancia del desarrollo infantil, en Chile han existido esfuerzo por medir dicho constructo a nivel población y longitudinal. Esto se refleja en la Encuesta Longitudinal de Primera Infancia (ELPI). Tal como se detalla:

*“ELPI corresponde a una encuesta de tipo longitudinal, diseñada para desarrollar el seguimiento periódico de una muestra de niños y niñas desde su nacimiento, con el propósito de conocer sus trayectorias de vida y los factores relevantes que explican su desarrollo.”* (Observatorio Ministerio Desarrollo Social, s.f.)

En las rondas de los años 2010, 2012 y 2017 de ELPI se utilizó el instrumento Inventario de Desarrollo Battelle para la medición de desarrollo infantil. (Centro Micro Datos, s.f.-a; Centro Micro Datos, s.f.-b; Observatorio Ministerio Desarrollo Social, s.f.)

A partir de estos antecedentes señalados, la importancia del desarrollo infantil hace necesaria su correcta interpretación. Pero un aspecto a considerar es que la medición de este constructo es compleja, en especial por su cambio a través del tiempo. A través de los datos poblacionales disponibles. ¿Es posible una interpretación longitudinal del desarrollo infantil en Chile, que de cuenta de su trayectoria?

## **Antecedentes Teóricos y Empíricos**

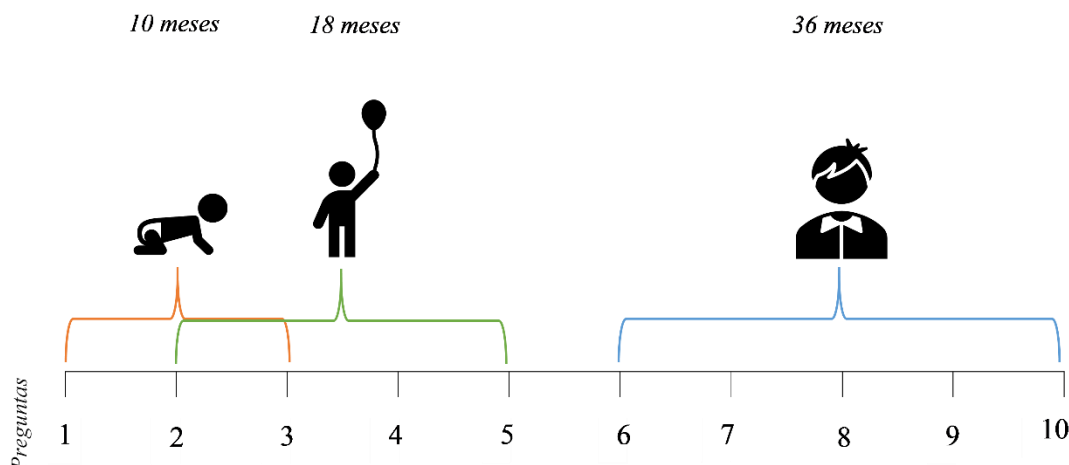
### ***Características de la Medición del Desarrollo Infantil***

La naturaleza cambiante en el tiempo del desarrollo infantil hace que cada una de las etapas de este se asocien a distintas conductas. Es decir, no se puede capturar la trayectoria del desarrollo bajo las mismas observaciones. Lo que se espera de un infante de 10 meses es distinto a lo que se espera de un infante de 3 años; a pesar de que en ambos casos nos estamos refiriendo a su desarrollo infantil.

Eso trae como consecuencia que, a la hora de medir este constructo, las preguntas a las que se enfrenta un infante cambian a medida que este crece. A modo de ejemplo se presenta el gráfico N°1.

## Gráfico N°1

### Representación de medición de Desarrollo Infantil



Se observa que se tienen 3 casos de 10, 18 y 36 meses respectivamente. Las preguntas y la cantidad de estas van a variar en cada uno de estos casos, debido a que se encuentran en un periodo del desarrollo cualitativamente distinto. Por lo tanto, si una medición de desarrollo cuenta con 10 preguntas, esto no implica que todos los casos respondan todo el instrumento, ni de la misma forma.

A pesar de esta disparidad de tareas a la que se ven expuesto, en los 3 casos se pretende medir el desarrollo infantil. La diferencia radica en que son distintos periodos dentro de la trayectoria del desarrollo.

### ***Interpretación de resultados sobre desarrollo infantil***

Considerando las complejidades que tiene la medición del desarrollo infantil, y en particular su cambio a través del tiempo, surge la pregunta de cómo interpretar los resultados que se han obtenido.

Si se tuviera un grupo de niñas y niños, de la misma edad y que se ven expuestos a las mismas preguntas, la interpretación de resultados pudiese basarse en la suma de sus éxitos y fracasos ante las tareas propuestas por el instrumento. Es decir, un puntaje total nos daría cuenta, eventualmente, de mayor o menor desarrollo entre cada niña y niño en ese momento.

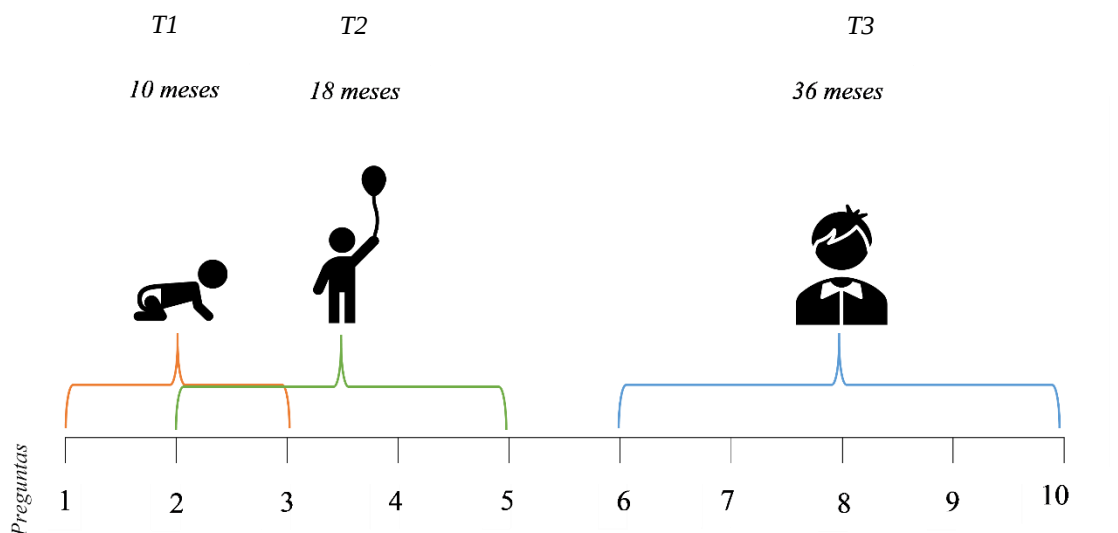
La limitación de este tipo de situaciones radica en que sólo son útiles para preguntas transversales, es decir, toda la información disponible se remite sólo a un periodo de tiempo.

Considerando que un aspecto fundamental para entender el desarrollo es que se comporta como una trayectoria, limitarse interpretaciones transversales no permite develar el crecimiento del desarrollo infantil en cada caso. Los puntajes totales, en el ejemplo recién expuesto, no reflejan un punto de inicio anterior que expliquen las diferencias de puntajes actuales.

A raíz de esto, se hace necesario tener interpretaciones de resultados que capturen el cambio continuo en el tiempo que tiene el desarrollo infantil, es decir, interpretaciones longitudinales. A modo de ejemplo se retoma el grafico N°1, pero ahora considerando que cada caso es el mismo individuo en tres periodos en el tiempo.

## Gráfico N°2

*Representación de medición de Desarrollo Infantil longitudinal*



Si se quisiera obtener los puntajes totales de cada periodo para compararlos entre sí, surge el problema detallado anteriormente. Los desafíos a los que se ve expuesto en cada periodo son distintos. Los puntajes totales de cada periodo del tiempo reflejan la suma de distintas preguntas y en diversa cantidad. Por lo tanto, los puntajes totales no son comparables. Ante esto surge el problema de cómo lograr interpretaciones longitudinales del desarrollo infantil.

Una de las alternativas es interpretar los resultados en base a norma. Según Thorndike, R.M & Thorndike, Christ (2010) en la interpretación en base a norma el puntaje se puede interpretar dentro de un marco normativo, al comparar el desempeño de una persona con el de otras. El desempeño es evaluado no en relación a una serie de tareas per se, sino en relación

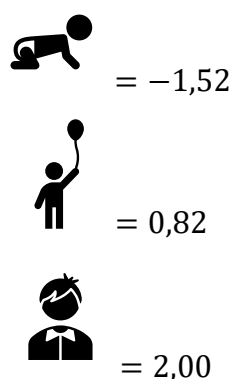
del desempeño de un grupo de referencia más general. La intención de este tipo de interpretación es saber si la persona se encuentra bajo el promedio, en el promedio o sobre el promedio en relación a un grupo de referencia. Con la interpretación bajo norma se busca uniformidad en el significado de test en test para comparar distintas pruebas, y uniformidad en la unidad de las escalas de puntajes en que se expresan los test

Una de las técnicas utilizadas en la interpretación de normas es la estandarización de puntajes. Al puntaje obtenido se le resta la media del grupo con el cual se compara y se divide por la desviación estándar (Thorndike, R.M & Thorndike, Christ, 2010)

Continuando con el ejemplo anterior, en cada uno de los periodos del caso se compara el resultado con cada grupo de referencia correspondiente. Es decir, en el periodo *T1* con un grupo de infantes de 10 meses, *T2* con un grupo de 18 meses y *T3* con un grupo de 36 meses. Dicho procedimiento da como resultado puntajes estandarizados comparables entre los 3 periodos. Tal como lo ilustra, a modo de ejemplo, en el gráfico N°3

### **Gráfico N°3**

*Representación puntajes estandarizados*



En el periodo 1 el caso se encontraba a 1,52 desviaciones estándar bajo el promedio del grupo, en el segundo periodo se encontraba a 0,82 desviaciones estándar sobre el promedio del grupo y en el tercer periodo se encontraba a 2 desviaciones estándar sobre el promedio del grupo. Considerando que en los puntajes estandarizados la diferencia de intervalos es la misma entre cada una de las comparaciones, se observa que el desarrollo infantil del caso ha ido en aumento, en relación a los grupos con los cuales se le compara en cada periodo. Por último, cabe recalcar que, al realizar una estandarización de puntajes, la

distribución queda fijada con una media de 0 y desviación estándar de 1 (Thorndike, R.M & Thorndike, Christ, 2010)

A pesar de aquello, la interpretación en base a norma, utilizando puntajes estandarizados según rangos de edad, implica ciertas limitaciones a la hora de tener interpretaciones longitudinales sobre el desarrollo infantil.

Para ilustrar este argumento se presenta la siguiente analogía. Para realizar evaluaciones nutricionales de niñas, niños y adolescentes se utilizan patrones de crecimiento (Ministerio de Salud, 2018). Dichos patrones de crecimiento se basan en estudios poblacionales para establecer normas, según edad y sexo, que permitan detectar el estado nutricional de cada caso. Por ejemplo, considerando el caso de la talla, se establece que  $<-2$  desviaciones estándar es talla baja. Además, se establecen señales de alertas, tales como: “Niña con pubertad iniciada que incrementa menos de 3 cm en 6 meses” (Ministerio de Salud, 2018, pág. 10); “Niña sin pubertad iniciada que incrementa menos de 2 cm en 6 meses” (Ministerio de Salud, 2018, pág. 10).

Considerando esta información se puede plantear el caso de una niña que a los 6 meses de edad su talla sea de 65 centímetros, con lo cual su desviación estándar es de 0. Luego, al año de edad su talla continúa siendo 65 centímetros, por ende, su desviación estándar es  $-2$ . Al año de edad sería considerada de talla baja, según la norma, y estaría dentro de una las señales de alerta por incrementar menos de 2 centímetros en 6 meses.

En el caso del desarrollo infantil, bajo la interpretación de normas, se podría acceder a la misma información. Es decir, una niña de 6 meses con 0 desviación estándar en una prueba de desarrollo infantil, y pasado 6 meses (al año de edad) obtiene  $-2$  desviaciones estándar en la misma prueba.

La diferencia entre ambas interpretaciones longitudinales es que la medición de la talla está expresada en una escala común, los centímetros. Para cada una de las edades, y por ende, de las etapas del desarrollo, la talla se expresa en la misma escala de centímetros. Al año de edad su desviación estándar era negativa ya que los infantes de un año miden más que los de 6 meses. Por ende, respecto a ese grupo se encuentra bajo la media; pero la medición de su talla se mantuvo estable entre ambos periodos. Siempre midió 65 centímetros.

En cuanto al desarrollo infantil, como se detalló anteriormente, las preguntas son distintas en cada una de las etapas del desarrollo, ya que los periodos son distintos y se

esperan observan conductas diversas. Con la interpretación bajo normas, en el ejemplo señalado, lo único que se puede saber es que a los 6 meses su desempeño era similar a su grupo de referencia y al año su desempeño fue menor en relación a otro grupo de referencia. No se puede saber qué aspectos del desarrollo cambiaron entre cada una de las mediciones. A diferencia de la talla, en que se conoce que la niña medía lo mismo entre las mediciones, en cuanto al desarrollo infantil no se sabe si su desarrollo es igual o distinto, o que aspectos es capaz de lograr en la primera medición en comparación con la segunda. Por lo tanto, la interpretación bajo normas del desarrollo infantil permite comparaciones con grupo de referencia, pero no entrega información sobre el crecimiento de la trayectoria del desarrollo infantil.

La falta de información se produce porque en la medición de desarrollo infantil no existe, de forma inmediata, una escala común. Con el afán de ejemplificar, se podría pensar en el mismo caso de la niña a los 6 meses de edad, pero ahora su talla se mide según las cuartas de la mano del profesional de salud “A”. Luego, al año de edad, se mide su talla pero ahora con las cuartas de la mano del profesional de salud “B”. A los 6 meses de edad la talla era 3 cuartas de la mano del profesional “A”, al año de edad la talla era 3,2 cuartas de la mano del profesional “B”. Al no estar en la misma escala de medición, no se puede argumentar que la niña creció 0,2 cuartas de la mano. La cuarta de la mano de cada profesional de salud es distinta. Igualmente se podría obtener un puntaje estandarizado considerando que al grupo de referencia se le midió en cada oportunidad con las mismas cuartas de cada profesional de salud. Por lo tanto, se podría obtener exactamente el mismo resultado, es decir, que al año de edad la niña está -2 desviaciones estándar en relación al grupo de referencia, pero no se podría saber en cuanto creció en su talla.

Este mismo ejemplo ilustra lo que ocurre en la medición de desarrollo infantil. Al ser distintas preguntas según la edad del caso, se pueden obtener puntajes totales al sumar los éxitos y fracasos en cada una de las preguntas. Pero estos puntajes totales no son comparables longitudinalmente, del mismo modo que se no podían comparar las cuartas de la mano de los distintos profesionales de salud. De igual forma, como se detalló, la interpretación bajo normas no permite tener información sobre la trayectoria del desarrollo.

Por lo tanto, se requiere que a partir de la diversidad de preguntas que miden desarrollo infantil, generar una escala común que permita dar cuenta del crecimiento de la trayectoria del desarrollo infantil.

### ***Escala común para el desarrollo infantil – Escalamiento Vertical***

Para la elaboración de una escala común que emerja desde las mediciones de desarrollo infantil es necesario realizar un procedimiento de escalamiento vertical.

El escalamiento vertical consiste en vincular pruebas que difieren en su dificultad, midiendo el mismo constructo, para colocarlas en una sólo escala común de puntuaciones (Kolen & Brennan, 2014; Harris, 2007). Este proceso busca respaldar inferencias validas en relación con el crecimiento de una variable a través del tiempo (Patz & Yao, 2007).

De esta definición se desprenden dos aspectos fundamentales del escalamiento vertical y su relación con el desarrollo infantil. En primer lugar, las pruebas difieren en su dificultad. Es decir, hay diversos niveles de dificultad dentro del instrumento de medición (Kolen & Brennan, 2014). Tal como se nombró anteriormente, las preguntas para capturar el desarrollo infantil en un caso de 6 meses son distintas a las de un caso de 3 años. Aplicar el mismo set de preguntas en ambos casos es inapropiado para la medición, debido a que resultarían ser muy avanzadas para el caso de 6 meses o muy elementales para el caso de 3 años.

En segundo lugar, el objetivo es tener puntuaciones de una variable que crece en el tiempo. El constructo que se mide debe ser el mismo en cualquier punto de la escala. Asumir que la diferencia de dos puntos de la escala es igual al crecimiento implica que cada una de las habilidades capturadas en algunos de estos puntos emanan del mismo constructo (Young & Tong, 2006). Tal como se mencionó anteriormente, el desarrollo infantil es una trayectoria. A medida que la o el infante crece es capaz de lograr tareas de mayor dificultad, apoyándose en los logros ya adquiridos. Con lo cual, este constructo representa un continuo en el tiempo.

Para la generación de una escala vertical no existe un procedimiento único, sino más bien se debe tomar una serie de decisiones en diversas partes del procedimiento. Además, cada una de estas decisiones conllevan a resultados distintos (Harris, 2007; Briggs & Weeks, 2009). Las decisiones por tomar son del orden de la estructura del test y, por otro lado, en relación al método estadístico a utilizar.

En cuanto a la estructura del test, este puede tener tres variantes. En primer lugar, diseño de ítem común. Al tratarse de test que aumentan en su dificultad dado ciertos niveles, esta estrategia busca tener ítems comunes entre niveles adyacentes. Tal como lo ilustra el gráfico N°4, los participantes responden ciertas preguntas según su edad (marcadas en color gris). El diseño genera elementos comunes entre los niveles adyacentes. Por ejemplo, los casos de 1 y 2 años responden la pregunta P2. Gracias a este diseño, los ítems comunes permiten vincular las puntuaciones de todos los niveles del test (Kolen & Brennan, 2014).

#### **Gráfico N°4**

*Representación Diseño Ítem común*

|        | P1 | P2 | P3 | P4 | P5 |
|--------|----|----|----|----|----|
| 1 año  |    |    |    |    |    |
| 2 años |    |    |    |    |    |
| 3 años |    |    |    |    |    |
| 4 años |    |    |    |    |    |

En segundo lugar, el diseño de grupo equivalente. Los examinados, según su nivel, son asignados aleatoriamente para rendir el test correspondiente a su nivel o el test correspondiente al nivel adyacente. Usualmente el nivel adyacente que se suele asignar es un nivel de menor dificultad. Los datos generados de esta manera sirven para colocar los puntajes de todos los niveles en una sola escala. Además, es importante destacar que con este diseño no es necesario que existan elementos comunes entre los niveles (Kolen & Brennan, 2014).

Por último, se encuentra el diseño de prueba de escalamiento. Se crea un test que abarca todos los niveles de dificultad, pero se aplica en una sola sesión. Por ende, todos los casos reciben el mismo test. Además, cada caso rinde el test correspondiente a su nivel. Los datos de cada nivel se vinculan con la prueba de escalamiento para generar una puntuación común (Kolen & Brennan, 2014).

Pasando al segundo eje de la toma de decisiones en el escalamiento vertical, se debe optar por el método estadístico. Existen los métodos Hieronymus, Thurstone e IRT (Kolen & Brennan, 2014). El presente trabajo se centrará en el último de estos.

Los modelos de teoría de respuesta al ítem (IRT) se enmarcan dentro de los modelos de variables latentes. Una variable latente es comprendida como una variable aleatoria no observada que es modelada a través de la variabilidad de respuestas observadas (Borsboom,

2005). En el contexto de los modelos IRT, el atributo que se desea medir es latente y se modela a partir de las respuestas que produce un individuo cuando se enfrenta a una tarea. Se puede observar si las personas responden correcta o incorrectamente preguntas, y a partir de ello estimar los valores de un atributo no observado asociado a cada persona y a cada ítem (Wilson, 2005; Thorndike & Thorndike, 2010).

Los modelos IRT descansan en cuatro supuestos. Son modelos probabilísticos, se asume la existencia de una variable cuantitativa subyacente, unidimensionalidad e independencia local. Los dos primeros son de carácter conceptual en relación al atributo a medir, mientras que los dos últimos son de naturaleza operativa (Torres Irribarra, 2018; Crocker & Algina, 2008).

El primer supuesto asume la existencia de un componente aleatorio. Por ende, son modelos probabilísticos, no deterministas. Se modela la probabilidad de observar un éxito o fracaso. Por otro lado, el segundo supuesto asume que lo que se desea medir puede ser modelado como una variable cuantitativa. Es decir, tanto las personas como los ítems poseen cierta cantidad o nivel de la variable latente (Torres Irribarra, 2018; Crocker & Algina, 2008).

En el caso del desarrollo infantil, se podría argumentar que ambos supuestos se cumplen. En primer lugar, existe un componente aleatorio emanado del muestreo de la población. En segundo lugar, se asume que el desarrollo infantil es una variable cuantitativa continua. De esto emana su característica de comportarse como una trayectoria en el tiempo.

En cuanto al tercer y cuarto supuesto, la unidimensionalidad establece que todas las correlaciones en las respuestas son debido a un único atributo cuantitativo subyacente. Se mide un solo atributo. Por otro lado, el supuesto de independencia local detalla que todas las respuestas serán independientes condicionadas al atributo latente. Es decir, todas las correlaciones observadas en los datos son consecuencia del atributo latente. Por ende, al condicionarlo por este se eliminará las correlaciones observadas (Torres Irribarra, 2018).

Ambos supuestos están relacionados, pero no son lo mismo. Si se cumple el supuesto de unidimensionalidad, necesariamente se cumple el supuesto de independencia local, pero lo contrario no es necesariamente cierto (Torres Irribarra, 2018).

En cuanto al escalamiento vertical, si se cumplen estos supuestos, los modelos IRT tienen la propiedad de invariancia de parámetros. Esto implica que los parámetros de los ítems son independientes de las características específicas de la muestra que se utiliza para

estimarlos. Por lo tanto, si examinados de distinto nivel responden las mismas preguntas, la dificultad de los ítems es la misma independiente del grupo que las responda (Briggs & Weeks, 2009).

Dentro del enfoque IRT, para el caso del escalamiento vertical, también se deben tomar decisiones en cuanto al modelo, el procedimiento de estimación y la estimación del parámetro de habilidad.

En primer lugar, los modelos dentro del enfoque IRT son el modelo de Rasch, 2pl y 3pl. El primero de ellos modela la probabilidad de que una persona responda correctamente a través de una función de la diferencia entre el efecto de la persona y el efecto del ítem, ambos en la misma escala. Este modelo considera que el desempeño de una persona en una pregunta es debido al “trade-off” entre la habilidad de la persona y la dificultad del ítem (Rasch, 1960); (Wilson, 2005). Una de las formas de representar esta relación es la siguiente:

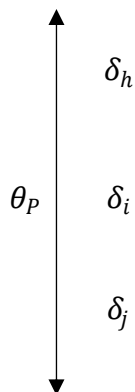
$$\text{Logit}(x_{ip} = 1|\theta) = \log \frac{Pr(x_{ip}=1)}{Pr(x_{ip}=0)} = \theta_p - \delta_i$$

Donde  $x_{ip} = 1$  representa a la de la persona  $p$  respondiendo correctamente el ítem  $i$ ,  $\theta_p$  representa la habilidad del atributo latente de la persona  $p$  y  $\delta_i$  representa la dificultad del ítem  $i$ . Por lo tanto, la relación establece que el logaritmo de las odds (logit) de responder correctamente de la persona  $p$  el ítem  $i$  según la variable latente es igual a la habilidad de la persona  $p$  menos la dificultad del ítem  $i$  (Torres Iribarra, 2018).

Como se nombró, en el modelo de Rasch, la habilidad de la persona y la dificultad del ítem están en la misma escala. Esto permite obtener interpretaciones sustantivas, ya que la habilidad obtenida se vincula a la probabilidad de responder correctamente los ítems del test (Wilson, 2005). Esto se ilustra en el gráfico N°5, donde la habilidad de la persona  $p$  se vincula a responder correctamente el ítem  $i$ . El ítem  $h$  es más difícil que el ítem  $i$ , y por ende sería menos probable que la persona  $p$  lo responda correctamente con su nivel de habilidad. Por último, el ítem  $j$  es más fácil que el ítem  $i$ , por lo tanto, la habilidad de la persona  $p$  es mayor del que se requiere para responder correctamente el ítem  $j$ .

### Gráfico N°5

#### *Habilidad y Dificultad en Modelo de Rasch*



A partir de esto, se puede ilustrar el siguiente ejemplo con relación al desarrollo infantil. Considerando el valor  $\theta_p = 2.0$  logit y los siguientes valores de los ítems:

**Tabla N°1**

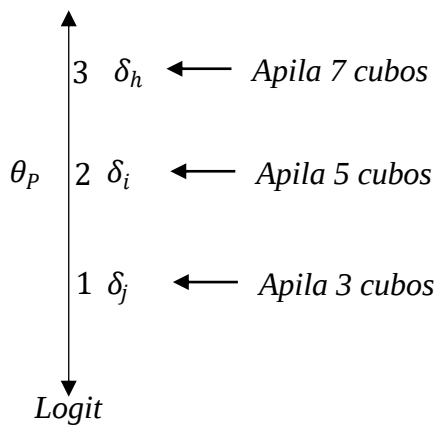
*Representación valores dificultad*

| Ítem | Descripción de la tarea | Valor (Logit) |
|------|-------------------------|---------------|
| j    | Apila 3 cubos           | 1.0           |
| i    | Apila 5 cubos           | 2.0           |
| h    | Apila 7 cubos           | 3.0           |

Se puede replicar el gráfico N°5 con los valores recién propuestos en la tabla N°1. Es así como, el examinado, dado su nivel de desarrollo, es probable que sea capaz de apilar 5 cubos. Con esto se puede establecer una interpretación sustantiva del constructo a medir.

## Gráfico N°6

### Valores de Habilidad y Dificultad en Modelo de Rasch



Por otro lado, los modelos 2pl y 3pl comparten la misma estructura que el modelo de Rasch, pero en cada uno de ellos se adicionan más parámetros. En el caso del modelo 2pl se agrega un parámetro de discriminación. El modelo de Rasch asume que todos los ítems se relacionan de la misma manera con la variable latente. En el modelo 2pl, con el parámetro de discriminación, se puede especificar diversos grados de relación entre cada uno de los ítems del instrumento y la variable latente, es decir, tener mayor o menor grado de relación (Leenen, 2014).

En el caso del modelo 3pl se adiciona un tercer parámetro de guessing (adivinabilidad). Con dicha incorporación se desea modelar los casos en que la obtención de una respuesta correcta se debe al azar. Por ejemplo, en una prueba de selección múltiple de cuatro alternativas hay un 25% de probabilidad de que la respuesta esté correcta (Leenen, 2014).

Los modelos 2pl y 3pl son más flexibles que el modelo de Rasch. Esto provoca que tengan un mejor ajuste a los datos. Pero la gran falencia de los modelos 2pl y 3pl, en comparación al modelo de Rasch, es que se pierde interpretabilidad. Es así como, en el modelo de Rasch se puede establecer que el ítem  $i$  es más fácil que el ítem  $h$  para todos los casos examinados. Con el modelo 2pl esta afirmación no es necesariamente correcta, ya que el parámetro de dificultad del ítem  $i$  pudiese ser menor que el del ítem  $h$ , pero a la vez ser más difícil que el ítem  $h$  para una porción de los casos por su grado de discriminación. De igual modo, en el caso del modelo 3pl, puede ocurrir que a pesar de que el ítem  $j$  tiene un

parámetro de dificultad menor que el ítem  $i$ , el primero tenga una adivinabilidad menor. Por ende, se vuelve más difícil obtener una respuesta correcta del ítem  $j$  en comparación al ítem  $i$  (Leenen, 2014).

La parsimonia del modelo de Rasch evita las situaciones descritas para los modelos 2pl y 3pl. Esto se debe a que el modelo de Rasch posee la propiedad de suficiencia estadística de sus parámetros. Es así como, el número total de respuestas correctas para un conjunto de personas ante un set de ítems constituye estadística suficiente para el parámetro de habilidad; de igual manera el número total de respuestas correctas para un set de ítems entre un conjunto de personas es estadística suficiente para el parámetro de dificultad. No existe información adicional para la estimación más que la suma de los marginales de los ítems y de las personas (Torres Iribarra, 2018). Como se detalló anteriormente, esto difiere en los modelos 2pl y 3pl, donde la adición de nuevos parámetros hace que la estimación deje de depender únicamente de la suma de los marginales.

En segundo lugar, en cuanto al procedimiento de estimación, este puede ser separado o concurrente. En el primero de ellos se hacen las estimaciones de parámetros para cada nivel de dificultad por separado. Luego las puntuaciones son vinculadas a una escala común por medio de transformaciones lineales. Por otro lado, en el procedimiento concurrente, todos los parámetros son estimados para la totalidad de grados de niveles de dificultad en un paso. (Briggs & Weeks, 2009). Toda la información disponible se utiliza para la estimación y no existen transformaciones posteriores (Kolen & Brennan, 2014).

En tercer lugar, para la estimación del parámetro de habilidad, usualmente se debe optar por los enfoques maximum likelihood (ML) o expected a posteriori (EAP). ML maximiza la probabilidad conjunta del patrón de respuesta de cada caso para determinar el nivel de habilidad más probable. Por otro lado, en el enfoque EAP, la distribución de probabilidad conjunta se pondera según la probabilidad de ciertos puntos de la distribución definida a priori para proporcionar una estimación de habilidad (Briggs & Weeks, 2009; Kolen & Brennan, 2014).

Las estimaciones de ML son asintóticamente insesgadas, pero ineficientes para casos cerca de las colas de distribución. Por otro lado, las estimaciones EAP reducen la varianza del error y poseen un menor error cuadrático medio, pero introducen sesgo en las estimaciones (Briggs & Weeks, 2009; Kolen & Brennan, 2014).

Cuando se comparan ambos estimadores, los resultados con EAP tienen la ventaja de que las estimaciones siempre existen, incluso con patrones de respuestas de sólo aciertos o sólo fallos. Además, las estimaciones con EAP son estables para toda la longitud del test (Kolen & Brennan, 2014).

Para la elección entre ML y EAP se debe considerar el “trade-off” entre eficiencia y sesgo de la estimación (Briggs & Weeks, 2009). Además, las estimaciones vía EAP consideran la distribución de puntajes de todos los examinados en el test. Por ende, la estimación de habilidad de un caso depende no sólo de las respuestas correctas que obtiene, sino también del desempeño de toda la población examinada (Kolen & Brennan, 2014).

Considerando todas las decisiones que se deben tomar al desarrollar el escalamiento vertical, es importante destacar que cada una de las opciones escogidas deben ser explícitas y según los propósitos que tengan los análisis (Harris, 2007; Briggs & Weeks, 2009).

### ***Mediciones de desarrollo infantil a nivel poblacional y longitudinal: El caso de Chile***

Como se mencionó, un aspecto clave para las mediciones de desarrollo infantil es contar con una recolección de datos de carácter poblacional y longitudinal.

Chile posee la Encuesta Longitudinal de Primera Infancia (ELPI). ELPI es un estudio elaborado por el Estado chileno con los objetivos de analizar el desarrollo infantil de niñas y niños, y caracterizar el entorno y los cuidadores de los infantes. Esta iniciativa es de carácter representativo de la población chilena y longitudinal (Observatorio Ministerio Desarrollo Social, s.f.)

A la fecha se cuenta con tres olas de ELPI realizadas los años 2010, 2012 y 2017. El año 2010 el total de evaluados fue de 14.161 entre las edades de 0 a 4 años. En la ronda 2012 el total de evaluados fue de 11.692 de la muestra de seguimiento más 2.746 casos de una muestra de refresco entre 0 y 3 años. Por último, en el año 2017 se evaluaron 11.421 de la muestra de seguimiento de las dos rondas anteriores más 4.406 casos de refresco de una muestra entre 0 y 5 años. (Centro Micro Datos, s.f.-c; Centro Micro Datos, s.f.-d; Centro UC Encuestas, 2018; Ministerio de Desarrollo Social, s.f.).

Las tres rondas de ELPI cuentan con el instrumento de desarrollo infantil multifacético Inventario de Desarrollo Battelle. En el momento de elaboración de esta tesis sólo se contaban con los datos disponibles de las rondas 2010 y 2012, por ende, se utiliza esa información.

**Inventario de Desarrollo Battelle.** El Inventario de Desarrollo Battelle (BDI) es un instrumento para medir desarrollo infantil que fue publicado en 1984 por Newborg. Cuenta con 341 ítems agrupados en las áreas personal social, adaptativa, motor, comunicación y cognitiva. El tiempo de aplicación varía entre 1 y 3 horas. Es apto para niñas y niños entre el nacimiento y 8 años de edad (Berls & McEwen, 1999).

Existen tres modalidades distintas para la producción de respuestas. En primer lugar, situaciones estructuradas en donde se especifican los procedimientos y materiales a utilizar para producir la respuesta deseada de la niña o niño. En segundo lugar, la observación de alguna conducta que emerja durante la evaluación. En tercer lugar, preguntas a la o el cuidador sobre ciertas conductas de la niña o niño (Newborg, Stock, & Wnek, 2011).

El sistema de valoración de los ítems es de 3 niveles con las categorías no logrado, logrado parcialmente y logrado. Es posible generar puntajes correspondientes a cada dimensión por separado y un puntaje global que caracteriza el desarrollo (Newborg, Stock, & Wnek, 2011).

Además, el Inventario de Desarrollo Battelle cuenta con una versión screening o abreviada diseñado para determinar, de forma expedita, qué áreas están deficitarias. Dichas áreas deficitarias son las que requerirían una evaluación con el instrumento completo. Para aquello, se determinan puntos de cortes según el número de desviaciones estándar que el examinador determina. Su duración oscila entre 10 y 30 minutos. El Inventario de Desarrollo Battelle screening cuenta con 96 ítems repartidos en 20 ítems en las áreas personal social, adaptativa y motora; y 18 en las áreas comunicación y cognitiva. Todos estos ítems son extraídos de la versión completa. Su elección se determinó por la correlación de las puntuaciones del ítem y la del área, y por su grado de dificultad (Newborg, Stock, & Wnek, 2011).

En cuanto a sus aspectos psicométricos el instrumento declara, por un lado, resultados satisfactorios de confiabilidad tanto para test-retest, como, inter-evaluador y error estándar de medición. Por otro lado, igualmente positivos, son los resultados de validez concurrente, de contenido, factorial, de constructo y predictiva (Berls & McEwen, 1999). En cuanto a estudios independientes, Snyder et al (1993), mediante un análisis factorial, encuentra estabilidad en el tiempo de las mediciones, pero alta correlación entre los dominios. De igual manera, Snyder et al (1993), por medio de la teoría de la generalibilidad, afirman

que la confiabilidad tiene resultados óptimos. Por otro lado, Jentzsch (1994) encuentra resultados favorables de la validez predictiva del BDI en una muestra de casos de necesidades especiales.

En relación al BDI Screening los autores no especifican las propiedades de validez y confiabilidad. En cuanto a estudios independientes, Glascoe & Byrne (1993), concluyen que la sensibilidad del instrumento es buena, pero su especificidad no es ideal. Por lo tanto, hay una sobre estimación de casos con retraso en el desarrollo para el grupo mayores de cuatro años y con cumpleaños reciente.

En 2005, Newborg publica una revisión del Inventario Battelle bajo el nombre de Inventario de Desarrollo Battelle 2ª edición (BDI-2). Los cambios incluyen reducir el número de subdominios, reposicionar ítems que estaban en subdominios eliminados, expandir las muestras normativas, revisión de pisos y techos e incorporación de nuevos ítems. El BDI-2 cuenta con 450 ítems con las mismas áreas e igual sistema de puntuación. También posee una versión screening compuesta por 100 ítems (Bliss, 2007).

Los autores declaran que la confiabilidad test-retest, consistencia interna y error estándar de medición tienen resultados favorables. Igualmente, positivos son la validez de constructo, de contenido y criterio (Bliss, 2007).

En cuanto a BDI-2 Screening, igual que su antecesor, no se declaran resultados específicos de validez y/o confiabilidad que emerjan de forma separada del BDI-2 (Bliss, 2007). En cuanto a investigaciones independientes, Elbaum, Gattamorta & Penfield (2010) hallaron, bajo el enfoque del Modelo de Rasch, resultados satisfactorios de confiabilidad. En relación a la validez se encontró resultados favorables en cuanto al ajuste de los ítems, el análisis de los componentes principales y la confirmación empírica del ordenamiento de los ítems. Cabe destacar que en esta investigación se utilizaron datos de una evaluación de BDI-2, y posteriormente se realizaron los análisis solo incluyendo los ítems correspondientes al Screening. No se realizó un levantamiento de datos utilizando únicamente el instrumento BDI-2 Screening.

El instrumento Inventario de Desarrollo Battelle, tanto en su primera como en su segunda versión, ha sido utilizado para responder preguntas de investigación longitudinales, tales como la relación entre los dominios del lenguaje, la cognición y socioemocional a través del tiempo (Gilliam & de Mesquita, 2000; Nip, Green, & Marx, 2011), la emergencia y

desarrollo de la comunicación verbal (Medeiros, Cress, & Lambert, 2016) o el efecto del consumo de cocaína en el embarazo sobre el desarrollo infantil posterior (Hurt, Malmud, Betancourt, Brodsky, & Giannetta, 2001). En todos estos casos, en que la pregunta de investigación se limita al crecimiento de la trayectoria del desarrollo infantil, se han utilizado puntajes estandarizados.

En el caso de Chile, el Inventario de Desarrollo Battelle fue analizado en el contexto de los requerimientos por parte de UNICEF, Ministerio de Desarrollo Social, Ministerio de Salud y Consejo de la Infancia para investigar los instrumentos de desarrollo infantil más eficaces. El ente ejecutor fue Medwave Estudios Limitada y llevo a cabo la tarea de evaluar casos entre 0 y 4 años con los instrumentos EEDP, TEPSI, BDI Screening y BDI. Cabe señalar que los resultados de Medwave Estudios Limitada sólo contemplan la creación de una base de datos con los resultados, no el análisis de estos. Para los objetivos de esta tesis, es relevante destacar que la investigación realizada por Medwave Estudios Limitada en el año 2017 declara: *“BDI-2 no está disponible en español en una versión validada para España y América Latina, por lo que no se comercializa legalmente en estas regiones”* (Medwave Estudios Limitada, 2017, pág. 25).

Para el caso de ELPI, el informe de resultados de evaluaciones de la primera ronda del año 2010 informa que el instrumento utilizado es BDI. Mientras que el informe de resultados de evaluaciones de la segunda ronda del año 2012 detalla que el instrumento utilizado es BDI-2 Screening (Centro Micro Datos, s.f.-a; Centro Micro Datos, s.f.-b). Es decir, entre ronda hay un cambio de versión y se pasa del instrumento completo al abreviado.

Sin embargo, a pesar de que el informe de resultados de evaluaciones de la segunda ronda del año 2012 de ELPI declara que el instrumento utilizado es BDI-2 Screening, las referencias señaladas en dicho documento son Newborg, Stock y Wnek del año 1996; siendo que BDI-2 se publica en el año 2005. Además, como detalló el informe de Medwave Estudios Limitada, no existen traducciones oficiales para España ni América Latina del BDI-2. Por ende, pareciese ser que el BDI-2 Screening ocupado en ELPI 2012 se basa en traducciones ad hoc del inglés al español para ser utilizado en Chile.

### **Planteamiento del problema**

A lo largo de esta tesis se ha establecido la importancia del desarrollo infantil, y, por ende, la necesidad de su medición e interpretación a escala población y longitudinal.

La característica de trayectoria a través del tiempo que posee el desarrollo infantil implica ciertos desafíos a la hora de su medición e interpretación.

Al ser una trayectoria, en cada etapa del desarrollo se esperan observar distintas conductas. Por lo tanto, a través de las mismas preguntas no se puede caracterizar la trayectoria del desarrollo infantil.

Como diversas preguntas caracterizan el mismo constructo, existe la imposibilidad de interpretar longitudinalmente el desarrollo infantil por medio de un puntaje total. La alternativa ante esto es interpretar los resultados bajo normas. Pero como se detalló, este tipo de interpretación no está en una escala común, por lo cual no entrega información sobre el crecimiento de la trayectoria del desarrollo.

Para lograr una escala común en el desarrollo infantil es necesario un procedimiento de escalamiento vertical. Este proceso implica decisiones respecto a la estructura del test y el procedimiento estadístico.

Con los datos disponibles en Chile a nivel población y longitudinal el desafío de esta tesis consiste en lograr interpretaciones longitudinales del desarrollo infantil. Esto por medio de un escalamiento vertical, considerando las distintas versiones y cambios en el instrumento Inventario del Desarrollo Battelle utilizado en ELPI. Con esto se pretende obtener información que dé cuenta de la trayectoria del desarrollo en las niñas y niños de Chile.

## **Objetivo**

Establecer una métrica común para la comparación de los resultados de Inventario del Desarrollo Battelle recolectados en ELPI que permita una interpretación longitudinal del desarrollo infantil de las niñas y niños en Chile

## **Metodología**

### ***Datos***

Los datos a utilizar corresponden a BDI de la ronda 2010 de ELPI y BDI-2 Screening de la ronda 2012 de ELPI.<sup>1</sup>

En cuanto al año 2010 los datos corresponden a 4.869 casos de entre 6 y 23 meses. De estos, 2.400 corresponden a mujeres. Por otro lado, en el año 2012, hay 14.087 casos de entre 6 y 83 meses. De los cuales, 6.952 corresponden a mujeres.

---

<sup>1</sup> Información entregada por la Subsecretaría de Evaluación Social vía ley de transparencia folio N°AI001T0001817

### ***Procedimiento Estadístico***

Cómo se detalló en la sección escala común para el desarrollo, en el escalamiento vertical se deben tomar decisiones para su realización.

**Decisión de la estructura del test.** En cuanto a la estructura del test se opta por el diseño de ítem común. Considerando los datos disponibles de las rondas de ELPI y que el Instrumento Inventario de Desarrollo Battelle fue aplicado en ambas rondas, se asume que existen ítems comunes.

La suposición de existencia de ítems comunes entre rondas debe analizarse con detenimiento debido a dos aspectos. En primer lugar, como se detalló, no se aplicó el mismo instrumento en ambas rondas. En la ronda 2010 se utilizó BDI, y en la ronda 2012 BDI-2 Screening. Por lo tanto, la existencia de ítems comunes depende de la relación entre BDI y BDI-2; y a la vez entre BDI-2 y BDI-2 Screening.

En segundo lugar, como se argumentó, se especula que se hicieron adaptaciones ad hoc para la aplicación del instrumento en la ronda 2012. Eventualmente este hecho podría dificultar la búsqueda de ítems comunes entre rondas.

A la luz de estos argumentos, se decide utilizar como insumo los manuales de evaluadores<sup>2</sup> de ambas rondas para indagar en los ítems comunes. Por lo tanto, para este paso no se hace uso del instrumento Inventario de Desarrollo Battelle comercializado.

Los ítems del instrumento inventario de Desarrollo Battelle, tanto en su versión completa como screening, se componen de tres elementos. En primer lugar, la conducta, que corresponde a la acción que se espera observar en el sujeto evaluado. En segundo lugar, el procedimiento, que consiste en la explicación de los procesos que debe realizar la evaluadora o evaluador para demostrar la existencia de la conducta. Por último, el criterio de puntuación, donde se detalla que aspectos, de los que realiza la evaluada o evaluado a raíz del procedimiento, corresponden a una puntuación 2, 1 o 0.

En una primera instancia el trabajo consistió en analizar la conducta de cada ítem, en cada una de las versiones de Battelle, y dejar registro de los ítems que eran altamente similares en cuando a la conducta. En esta instancia se permitía que un ítem estuviese pareado con más

---

<sup>2</sup> Información entregada por la Subsecretaría de Evaluación Social vía ley de transparencia folio N°AI001T0001678

de uno. Por ejemplo, que una pregunta del BDI tuviera dos posibles candidatos de ítems comunes en el BDI-2 Screening, o viceversa. Un ejemplo de esto es el siguiente caso:

**Tabla N°2**

*Ejemplo ítems con dos posibles candidatos*

| <b>Tipo</b>        |               |                                                                                                                                           |
|--------------------|---------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Battelle</b>    | <b>N°Item</b> | <b>Conducta</b>                                                                                                                           |
| BDI-2<br>Screening | 51            | El niño/a responde a las preguntas de qué y quién.                                                                                        |
| BDI-2<br>Screening | 53            | El niño/a responde a las preguntas de dónde y cuándo.<br><br>Responde a preguntas que incluyen las palabras: “qué, quién, dónde y cuándo” |
| BDI                | CM 13         |                                                                                                                                           |

En una segunda instancia de revisión se realizaron dos procesos. En primer lugar, se examinaron nuevamente los ítems propuestos como comunes en la fase anterior, ahora considerando todos los aspectos del ítem, es decir, la conducta, el procedimiento y criterio de puntuación. Esto para concluir si efectivamente corresponde a un ítem común o no teniendo en cuenta toda la complejidad del ítem. En segundo lugar, en los casos en que algún ítem estaba pareado con más de uno, se seleccionaba solo uno.

En esta segunda instancia de revisión, determinar si un par de ítems eran comunes o no pasaba por los procesos ya descritos. Siempre amparando que, en la aplicación final de cada una de esas preguntas, la situación a la que se vio expuesta el infante fuese similar.

La conclusión de esta revisión fue el hallazgo de 46 pares de ítems comunes entre las versiones BDI y BDI-2 Screening en las rondas ELPI 2010 y 2012.

Considerando estos 46 pares de ítems, se procedió a una tercera fase. Esta consistió en analizar si existían alguna diferencia entre estos ítems considerando la conducta, el procedimiento y los criterios de puntuación. Cabe consignar que estos 46 pares de ítems comunes son altamente similares; pero el objetivo de esta tercera fase es consignar si en cada uno de los tres elementos que componen el ítem (conducta, procedimiento y criterio de puntuación) estos son idénticos.

Los resultados de esta fase indicaron que hay ítems comunes que difieren en la conducta, procedimiento y/o criterios de puntuación. A continuación, a manera de ejemplo, se muestran tres casos en que el par de ítem comunes difieren en sólo en un aspecto.

**Tabla N°3**

*Diferencia en la Conducta*

| <b>Tipo Battelle</b> | <b>N° Item</b> | <b>Conducta</b>                                                | <b>Procedimiento</b>                                                                                                                                                                                                                                    | <b>Criterio de puntuación</b>                                                                                                                                                                                                     |
|----------------------|----------------|----------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BDI-2 Screening      | 1              | <b>El niño/a succiona con movimientos suaves y coordinados</b> | Pregúntele al cuidador: “¿El niño/a chupa bien la mamadera o el pecho? ¿Cuánta leche o relleno bota cuando toma leche? ¿Cuánta leche o relleno bota cuando toma leche? ¿El niño se trapica cuando toma leche? Si lo hace: ¿Qué tan seguido ocurre esto? | El niño succiona con movimientos suaves y coordinados; derrama, tose y se ahoga mínimamente: 2: Normalmente 1: A veces 0: Rara vez o nunca                                                                                        |
|                      |                | <b>Bebe de una taza con ayuda</b>                              | Se pregunta a la madre o tutor(a) si el niño bebe en una taza, cuánta ayuda necesita, cuánta cantidad derrama y cuántas veces lo hace.                                                                                                                  | El niño bebe en una taza con ayuda. También se puntúa si el niño bebe sin ayuda. Se permite que derrame algo 2 puntos: Normalmente (90% de las veces o más) 1 punto: A veces (50% de las veces) 0 puntos: Muy pocas veces o nunca |

La tabla N°3 muestra que la conducta en cada uno de los ítems es similar, pero no idéntica. En el caso de BDI-2 Screening se pretende develar como se alimenta de líquidos el infante de una mamadera o pecho. Mientras que en el caso del BDI también está el objetivo de indagar como se alimenta de líquidos el infante; pero en este caso hay una referencia explícita a la taza y no a una mamadera o pecho. Como se observa, los procedimientos y criterios de puntuación son altamente similares. El hecho de que la conducta haga referencia

a situaciones distintas de alimentación de líquidos hace que este par de ítem comunes no sean idénticos.

**Tabla N°4**

*Diferencia en el Procedimiento*

| <b>Tipo Battelle</b> | <b>N° Item</b> | <b>Conducta</b>                      | <b>Procedimiento</b>                                                                                                                                                                                                                                   | <b>Criterio de puntuación</b>                                                                                                                                       |
|----------------------|----------------|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BDI-2 Screening      | 69             | El niño/a corre 3 metros sin caerse. | <b>Pregunta: Pregunte al cuidador:</b><br><b>“¿EL NIÑO/A ES CAPAZ DE CORRER AL MENOS 3 METROS SIN CAERSE?”.</b>                                                                                                                                        | El niño/a corre sin caerse: 2: Al menos 3 metros. 1: Entre 1.5 y 3 metros. 0: Menos de 1.5 metros.                                                                  |
|                      |                |                                      | <b>Se pide a la madre o tutor(a) que se sitúe en un lugar determinado de la habitación. Se lleva al niño a unos 3,5 m de distancia respecto al adulto y se le dice: “Quiero ver como corres; ve corriendo hasta donde está... (mamá, tutor, etc.)”</b> | El niño corre como mínimo 3m sin caerse. No se puntúa si el niño anda deprisa. 2 puntos: Corre 3 m o más 1 punto: Corre entre 2,9 m y 1,5m 0 puntos: menos de 1,5 m |
| BDI                  | M 40           | Corre 3 m sin caerse                 |                                                                                                                                                                                                                                                        |                                                                                                                                                                     |

En la tabla N°4 muestra como ambas preguntas tienen exactamente la misma conducta y criterios de puntuación. En ambas situaciones se desea que el caso demuestre la habilidad motora de correr tres metros sin caerse, y puntuar dicha conducta con 2,1 o 0 según cierto parámetro de distancia. La diferencia radica en que el procedimiento de ambas preguntas es distinto. En el caso del BDI-2 Screening la evaluadora o evaluador le pregunta al cuidador si dicha conducta ocurre, mientras que en el BDI la evaluadora o evaluador realiza una actividad para observar si el infante es capaz de correr tres metros sin caerse. Ambos ítems se tornan distintos considerando esta diferencia de procedimiento. Que el cuidador recuerde situaciones en donde estime que la niña o niño camina más de tres metros sin caerse es incomparable a realizar una actividad en donde la evaluadora o evaluador juzgue la capacidad motriz del infante al correr tres metros sin caerse.

**Tabla N°5**

*Diferencia en Criterios de puntuación*

| <b>Tipo Battelle</b> | <b>N° Item</b> | <b>Conducta</b>                                                                                               | <b>Procedimiento</b>                                                                                                                                                                          | <b>Criterio de puntuación</b>                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------|----------------|---------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BDI-2 Screening      | 62             | El niño/a mantiene las palmas de las manos abiertas y los puños relajados cuando no está sujetando un objeto. | Obs: Durante la evaluación, observar las manos del niño/a cuando no esté sosteniendo un objeto, para determinar si las manos están en una posición abierta, relajadas.                        | <b>El niño/a mantiene las palmas de las manos abiertas y los puños relajados cuando no está sujetando un objeto. 2: Normalmente. 1: A veces. 0: Rara vez o nunca.</b>                                                                                                                                                                                                                                                                           |
| BDI                  | M 45           | Mantiene las manos predominantemente abiertas                                                                 | Se observan las manos del niño cuando no está sujetando ningún objeto para determinar si tiene las manos abiertas (en oposición a los puños cerrados, característicos de los recién nacidos). | <b>Normalmente el niño mantiene las manos abiertas cuando no sujeta ningún objeto. Se acepta que el niño cierre las manos por unos segundos. Considerar como tiempo de observación 1 minuto. 2 puntos: El niño pasa más de 50 segundos predominantemente con las manos abiertas. 1 punto: El niño pasa entre 30 y 49 segundos predominantemente con las manos abiertas, 0 puntos: El niño pasa menos de 29 segundos con las manos abiertas.</b> |

Las preguntas de la tabla N°5 buscan evaluar la capacidad motora del infante a través de la mantención abierta de sus palmas. Además, en ambas preguntas es la evaluadora o evaluador, que, por medio de la observación, debe determinar si dicha conducta existe o no en el caso evaluado. Pero los criterios de puntuación difieren notoriamente. Mientras que en la pregunta M 45 del BDI las puntuaciones 2,1 y 0 están definidas por tramos de segundos;

en el caso del BDI-2 Screening las puntuaciones no están definidas según algún intervalo de tiempo; sino por el criterio de la evaluadora o evaluador de lo que corresponde “normalmente”, “a veces” o “rara vez o nunca”.

En los tres ejemplos recién señalados, algunos de los tres aspectos descritos provocan que los ítems no sean exactamente idénticos. Lo problemático de, a pesar de estas diferencias, considerarlos como idénticos, es que la niña o niño no se vio expuesto a la misma situación a la hora de evaluar su desarrollo infantil.

En el primer caso no es posible asegurar que para el mismo infante es igual beber de una mamadera que de una taza. Se podría argumentar que es “más fácil” beber de una mamadera que de una taza.

En el segundo caso no es comparable preguntarle a la cuidadora o cuidador si la niña o niño corrió tres metros sin caerse, en lugar de realizar una actividad donde se demuestre dicha conducta. Es plausible pensar que es “más fácil” demostrar que se puede correr tres metros sin caerse no realizando la actividad y preguntándole a la cuidadora o cuidador.

Finalmente, en el tercer ejemplo, los criterios de puntuación difieren ya que en el caso del BDI-2 Screening no es explícito un rango de tiempo para puntuar. No es posible asegurar si los ítems sean idénticos. Eventualmente en algunos casos sería “más fácil” y en otros “más difícil” que la evaluadora o evaluador considere un rango de tiempo para puntuar el ítem, en contraposición a que no hubiera ningún rango. Los criterios difieren en el grado de apreciación de la persona que aplica el test.

Considerando estas diferencias que existían en conducta, procedimiento y/o criterio de aprobación en los ítems comunes; es que se procedió a seleccionar los pares de ítems que no tuvieran diferencia en ninguno de estos tres aspectos. Dicha selección dejó como resultado un total de 20 pares de ítems comunes. Un ejemplo de estos ítems comunes es el siguiente:

## Gráfico N°7

### Ejemplo Ítem común

| Tipo Battelle   | N°Item | Conducta                                          | Procedimiento                                                                                                                                                                                                                                                                                                                                                                                                                                              | Criterio de puntuación                                                                                                                                                                                                                          |
|-----------------|--------|---------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BDI-2 Screening | 64     | El niño/a cambia un objeto de una mano a la otra. | Tarea: Coloque un cubo en la mano del niño/a y luego preséntele un segundo cubo al niño/a frente a su misma mano. El niño/a debe pasar el primer cubo a la otra mano para poder tomar el segundo cubo. Si el niño/a transfiere correctamente el cubo, preséntele un tercer cubo para fomentar la transferencia a la otra mano. Repita este procedimiento una vez si el niño/a no logra hacer 2 transferencias al primer intento.                           | El niño/a cambia un objeto de una mano a la otra:<br>2: 2 veces.<br>1: 1 vez.<br>0: Ninguna.<br>Nota: El niño/a NO puede usar su boca, el cuerpo, o una superficie, tal como la mesa, para ayudar a transferir los cubos de una mano a la otra. |
| BDI             | M 48   | Se pasa un objeto de una mano a otra              | Se coloca al niño de espalda o sentado (con el apoyo que necesite). Se le pone un bloque en la mano y luego se le presenta un segundo bloque en la misma mano que tiene el anterior. El niño debe pasarse el primer bloque a la otra mano, para poder tomar el segundo. Si hace el cambio se le presenta un tercer bloque para animarle a hacer otro cambio. Puede repetirse el procedimiento si el niño no realiza los dos cambios en la misma aplicación | El niño se pasa el bloque de una mano a la otra dos veces. No se permite que utilice la boca, el cuerpo o la mesa.<br>2 puntos: 2 cambios, considerando ambos intentos.<br>1 punto: 1 cambio<br>0 puntos: Ningún intento                        |

**Decisiones del modelo estadístico.** En cuanto al modelo estadístico se decide por escoger una perspectiva IRT debido a su característica de invariabilidad de los parámetros, aspecto clave en el escalamiento vertical.

El modelo escogido, dentro de esta perspectiva, es el modelo de Rasch. Teniendo en cuenta que el objetivo de esta tesis es lograr interpretaciones longitudinales sobre el desarrollo infantil, el modelo de Rasch se alza como una alternativa idónea por su capacidad de generar interpretaciones sustantivas. A su vez, considerando que los ítems del Inventario del Desarrollo Battelle son politómicos se optan por usar un modelo de crédito parcial. Este modelo es una alternativa para los casos en que las respuestas son variables ordenadas (Masters, 1982); (Wu & Adams, 2007); (Rost, 2012).

En relación a la estimación se opta por una concurrente en lugar de separada. Debido a que la estimación concurrente ocupa toda la información disponible, se espera que los resultados sean más estables. Sin embargo, tiene la desventaja de que es más difícil saber si algún ítem se comporta distinto a lo largo de los diversos niveles de la prueba (Kolen & Brennan, 2014). Debido a esto, más adelante se presentan análisis psicométricos de los ítems comunes para resguardar las interpretaciones de los análisis.

Por último, en cuanto al estimador de habilidad, se opta por expected a posteriori (EAP).

Para los análisis estadísticos se utiliza el programa R y el paquete TAM. La estimación de los parámetros en dicho paquete es por maximum likelihood (ML) y el algoritmo expectation maximization (EM). Debido a esto, se pueden estimar los parámetros a pesar de que existan datos perdidos en las preguntas de los test. No hay necesidad de técnicas de imputación. El único requisito es que no existan casos en que todo el test sean datos perdidos. Este tipo de situaciones sólo ocurrían en el año 2012, y se sacaron del análisis 300 casos.

**Análisis psicométricos ítems comunes.** A continuación, se presentan análisis psicométricos de los ítems comunes propuestos. Este paso es fundamental debido a dos aspectos. En primer lugar, la estimación será concurrente, por ende, no se podrá saber si hay ítems que se comporten de manera diferente a medida que la edad cambia. En segundo lugar, los ítems comunes que se proponen tendrán la función de anclar ambas pruebas, es decir, serán capaces de proveer información de cada una de estas.

En los procesos descritos en la sección identificación de ítems comunes, es posible afirmar que conceptualmente ambos pares de ítems son idénticos, por ende, fueron aplicados

en cada versión de Battelle. Pero, a pesar de aquello, puede ocurrir que estos ítems no posean adecuadas propiedades psicométricas. Si ese fuese el caso, la información que son capaces de proveer en cada una de las pruebas no son comparables.

Para estos propósitos se llevan a cabos dos análisis para los ítems comunes propuestos. En primer lugar, revisar el ajuste y, en segundo lugar, analizar la consistencia de la dificultad a lo largo de las edades de los casos en que fueron aplicados en cada prueba.

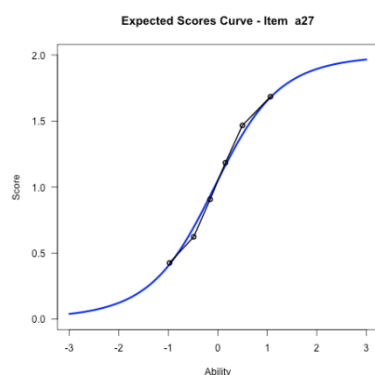
Cabe destacar que los análisis psicométricos que se detallaran fueron realizados considerando cada aplicación de Battelle por separado y por medio del modelo de crédito parcial.

En primer lugar, se detalla el ajuste de los ítems. En la teoría IRT, la curva característica del ítem (CCI), detalla la relación entre la variable latente y las respuestas observadas. La teoría asume que esta relación forma una curva sigmoide que representa una relación monótonamente creciente. Es decir, es más probable responder correctamente una pregunta cuando la persona está más alto en el atributo latente, o la pregunta está más abajo (Wilson, 2005).

Como ejemplo, los gráficos N°7 y N°8 muestran un correcto ajuste de ítems del BDI y BDI-2 Screening, respectivamente.

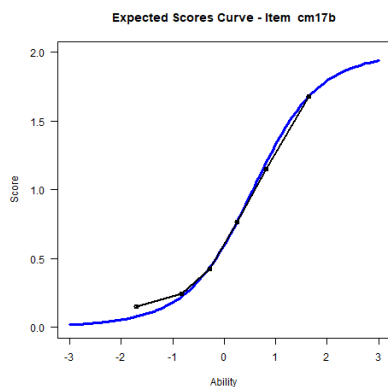
### **Gráfico N°8**

*Ejemplo ajuste correcto BDI*



## Gráfico N°9

### *Ejemplo ajuste correcto BDI-2 Screening*

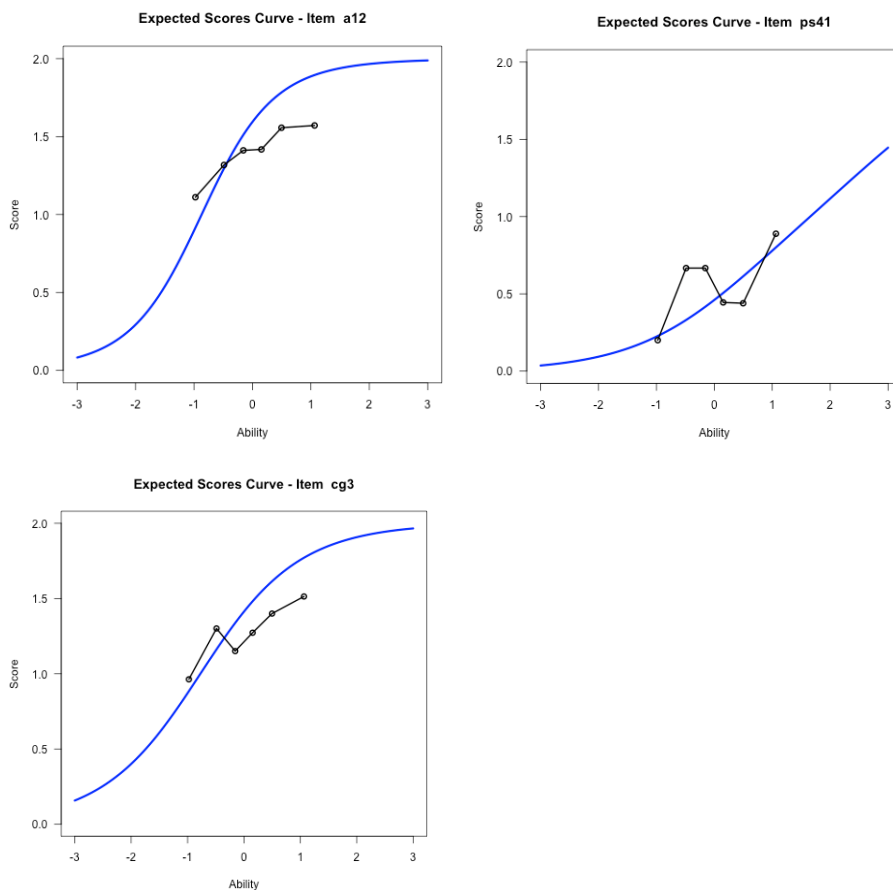


En la medida que los datos no se ajustan a esta forma en la relación entre la variable latente y las respuestas observadas, las conclusiones de dichos ítems no son justificadas (Wilson, 2005).

Considerando esto, los ítems comunes del BDI que no poseían un ajuste adecuado fueron los siguientes:

## Gráfico N°10

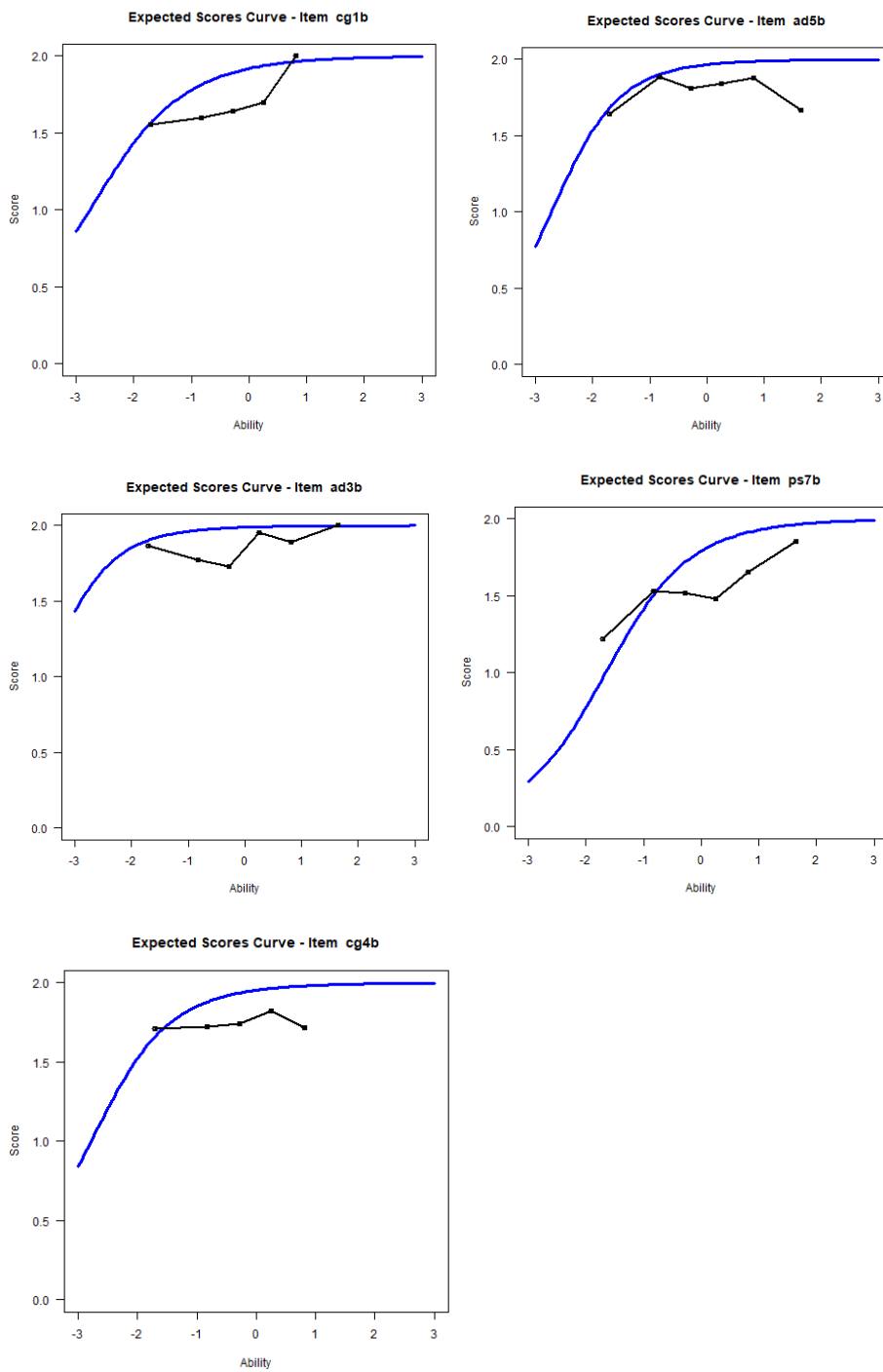
### Ítems con ajuste no adecuado en BDI



En cuanto en el BDI-2 Screening, las preguntas que no poseían un ajuste correcto fueron las siguientes:

## Gráfico N°11

*Ítems con ajuste no adecuado en BDI-2 Screening*



El segundo análisis psicométrico para los ítems comunes fue analizar su estabilidad a través de las distintas edades. Como se nombró anteriormente, una de las propiedades de los modelos IRT es la invarianza de los parámetros. Considerando esto, se opta por comparar el parámetro de dificultad entre los tramos de edad adyacentes que se proponen en cada una de las versiones de Battelle.

Para esto se estima, para cada uno de los tramos de edad por separado, el parámetro de dificultad de los ítems. Luego se comparan los parámetros de dificultad entre tramos de edades adyacentes en un gráfico de dispersión.

Se espera que en los gráficos de dispersión las dificultades de los ítems formen una diagonal. Si alguno de los ítems comunes propuestos se dispersa más allá de la diagonal, daría cuenta de que no es estable a través del tiempo.

A partir de los gráficos del anexo N°1 para el caso del BDI y del anexo N°2 en el caso del BDI-2 Screening, se concluye que no existen ítems comunes que sean inestables a través de los tramos de edad.

Considerando todos los análisis previos, tanto los que permitieron dilucidar que ítems eran comunes, como los que permitieron ver las propiedades psicométricas de estos; se proponen un anclado con 14 ítems comunes. (Ver anexo N°3).

## **Resultados**

### ***Comparación de estimaciones ancladas y libres***

Con los ítems comunes ya definidos para el anclaje, se procede a estimar los parámetros del modelo de Rasch con todos los ítems y todos los casos, tanto del año 2010 como del año 2012 en un solo paso (de forma concurrente). Por lo tanto, se utilizan 278 ítems, de los cuales 14 están anclados, es decir, 14 son idénticos entre las bases 2010 y 2012. Los resultados obtenidos de dicho procedimiento se comparan con las estimaciones de los parámetros de cada uno de los años estimadas de forma libre, sin ítem anclados (192 ítems del BDI del año 2010 y 100 ítems del BDI-2 Screening del año 2012).

Para comparar esos resultados se presentan los siguientes gráficos que muestran las dificultades para la versión con anclaje versus sin anclaje para cada año. Se etiquetan sólo los ítems anclados para mayor claridad.

### Comparación versiones con y sin anclaje en BDI 2010



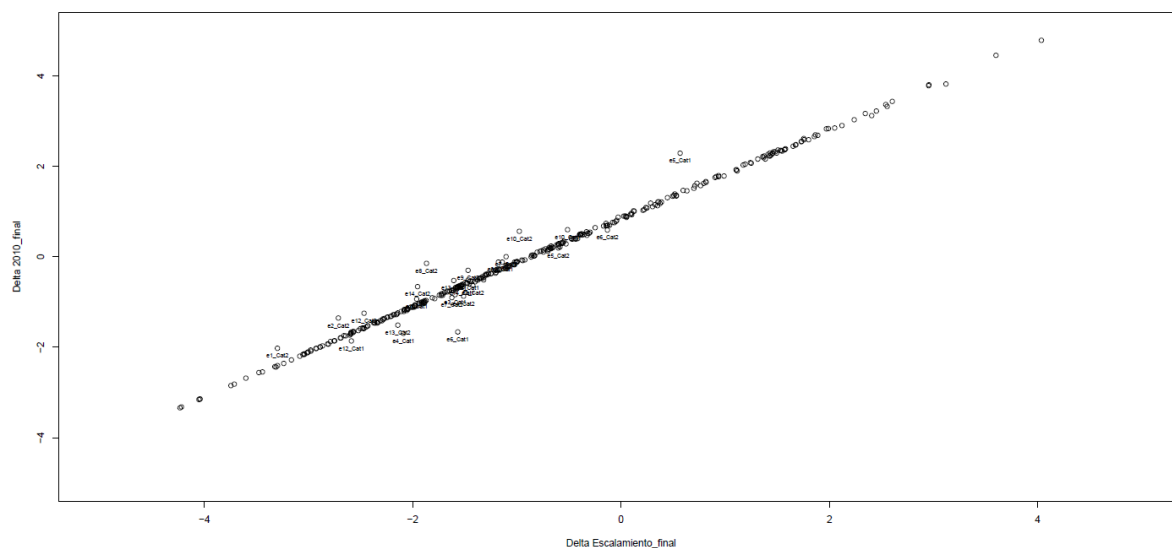
36

ítems. En cuanto al grafico N°13 que representa el año 2012 no se observan dispersiones lejos de la diagonal con una correlación de 0,9516 de los ítems comunes.

A raíz de los resultados mostrados se toma la decisión de quitar los pares de ítems ancla e11 (m56 del BDI y mt14b del BDI-2 Screening) y proponer una solución final de 13 ítems anclados. A continuación, se muestran los mismos gráficos recientemente presentados con la modificación propuesta.

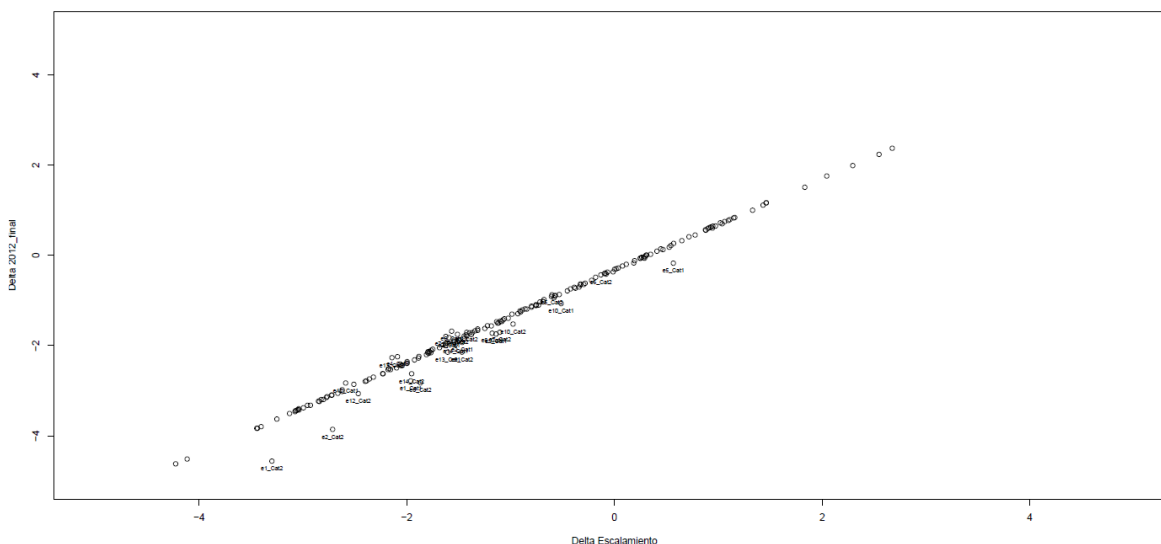
#### Gráfico N°14

*Comparación versiones con y sin anclaje en BDI 2010 versión final*



## Gráfico N°15

*Comparación versiones con y sin anclaje en BDI-2 Screening 2012 versión final*



Se observan en ambos gráficos que los ítems comunes están cercanos a la diagonal, esto quiere decir que en ninguno de los dos casos existen situaciones en que un mismo ítem se vuelva más difícil o fácil cuando se compara las estimaciones libres y las estimaciones con ítems anclados.

### ***Comparación de estimaciones entre versiones de anclaje***

A pesar de los buenos resultados del anclaje propuesto, la cantidad de ítems anclados ( $N^{\circ}=13$ ) representa un bajo porcentaje de los ítems totales de cada versión (6,8% para BDI y 13% para BDI-2 Screening). A raíz de aquello, se proponen 2 soluciones adicionales con mayor número de ítems ancla.

La versión 2 de ítems anclados corresponden a los 20 ítems comunes seleccionados previos a los análisis psicométricos. Son los 13 ítems pares de ítems de la solución final, más el par de ítem e11, más 6 pares ítems descartados por no presentar un ajuste adecuado. La versión 3 de ítems anclados corresponden a los mismos 20 ítems ya explicados en la versión 2, más 7 ítems que se agregaron, relajando algún criterio de conducta, procedimiento o puntuación, para lograr tener al menos 5 ítems de cada área (Ver anexo N°4).

En la versión 2 los ítems anclados representan un porcentaje 10,4% para BDI y 20% para el BDI-2 Screening. En el caso de la versión 3 representan un porcentaje de 14,1% y 27%, para las versiones BDI y BDI-2 Screening respectivamente.

Considerando estas versiones adicionales se procede a comparar la versión 1 (V1) propuesta inicialmente con 14 ítems anclados, la versión final que emergió de la versión 1, las versiones 2 y 3 recientemente explicadas; y las estimaciones originales de cada año estimadas de forma libre y sin ítems anclados.

Para las comparaciones se utilizará tanto el parámetro de dificultad como el de habilidad. En el caso del parámetro de dificultad, al igual que se hizo anteriormente, se comparan sólo las dificultades de los ítems anclados en cada caso. Por otro lado, en el parámetro de habilidad, en todas las versiones se comparan todos los casos de cada año.

**Tabla N°6**

*Correlaciones parámetros dificultad ítems ancla BDI año 2010*

|          | ORIGINAL | FINAL  | V1     | V2     | V3     |
|----------|----------|--------|--------|--------|--------|
| ORIGINAL | 1        | 0,9085 | 0,9064 | 0,9058 | 0,9040 |
| FINAL    | 0,9085   | 1      | 0,9997 | 0,9996 | 0,9986 |
| V1       | 0,9064   | 0,9997 | 1      | 0,9999 | 0,9996 |
| V2       | 0,9058   | 0,9996 | 0,9999 | 1      | 0,9997 |
| V3       | 0,904    | 0,9986 | 0,9996 | 0,9997 | 1      |

**Tabla N°7**

*Correlaciones parámetros dificultad ítems ancla BDI-2 Screening año 2012*

|          | ORIGINAL | FINAL  | V1     | V2     | V3     |
|----------|----------|--------|--------|--------|--------|
| ORIGINAL | 1        | 0,9502 | 0,9516 | 0,9519 | 0,9521 |
| FINAL    | 0,9502   | 1      | 0,9997 | 0,9996 | 0,9986 |
| V1       | 0,9516   | 0,9997 | 1      | 0,9999 | 0,9996 |
| V2       | 0,9519   | 0,9996 | 0,9999 | 1      | 0,9997 |
| V3       | 0,9521   | 0,9986 | 0,9996 | 0,9997 | 1      |

**Tabla N°8***Correlaciones parámetros habilidad BDI año 2010*

|          | ORIGINAL | FINAL  | V1     | V2     | V3     |
|----------|----------|--------|--------|--------|--------|
| ORIGINAL | 1        | 0,9504 | 0,9649 | 0,9713 | 0,9956 |
| FINAL    | 0,9504   | 1      | 0,9678 | 0,9624 | 0,9509 |
| V1       | 0,9649   | 0,9678 | 1      | 0,9987 | 0,9774 |
| V2       | 0,9713   | 0,9624 | 0,9987 | 1      | 0,9841 |
| V3       | 0,9956   | 0,9509 | 0,9774 | 0,9841 | 1      |

**Tabla N°9***Correlaciones parámetros habilidad BDI-2 Screening año 2012*

|          | ORIGINAL | FINAL  | V1     | V2     | V3     |
|----------|----------|--------|--------|--------|--------|
| ORIGINAL | 1        | 0,993  | 0,9931 | 0,9933 | 0,9949 |
| FINAL    | 0,993    | 1      | 0,9999 | 0,9999 | 0,9989 |
| V1       | 0,9931   | 0,9999 | 1      | 0,9999 | 0,999  |
| V2       | 0,9933   | 0,9999 | 0,9999 | 1      | 0,9993 |
| V3       | 0,9949   | 0,9989 | 0,999  | 0,9993 | 1      |

En cuanto a la dificultad, se observa como los resultados de la versión original se correlacionan con las demás versiones en un 0,90 y 0,95, para los casos de BDI y BDI-2 Screening respectivamente. En cuanto al anclaje final, se correlaciona en un 0,99 con los demás anclajes propuestos, tanto para el caso de BDI como BDI-2 Screening.

Respecto a la habilidad, para BDI, la versión original se correlaciona en 0,95 con el anclaje final y aumenta progresivamente hasta 0,99 en la versión 3 del anclaje. El anclaje final se correlaciona entre un 0,95 y 0,96 con las otras versiones. Para el caso de BDI-2 Screening, todas las versiones de anclaje y la habilidad original se correlacionan un 0,99 entre ellas.

Estos hallazgos indican que a pesar de que la solución final de anclaje propuesta sólo contempla 13 ítems, los resultados de las estimaciones con otras versiones y con la estimación libre original no difieren en demasía y tiende a mantenerse estable. Por ende, a pesar de que las otras versiones propuestas consideran mayor información (mayor número de ítem anclados) estos no provocan una diferencia considerable en la magnitud de las estimaciones.

Es así como, adquiere relevancia el anclaje final propuesto, ya que los ítems ancla fueron seleccionados por ser comunes entre ambas las versiones del instrumento Battelle y por su correcto desempeño psicométrico.

## **Conclusión**

Con este trabajo se puede concluir que es posible realizar el escalamiento vertical considerando un diseño de ítem común, un procedimiento estadístico en base al modelo de Rasch y el modelo de crédito parcial, una estimación concurrente y un estimador de habilidad en base a EAP; utilizando los datos disponibles del instrumento Battelle en las rondas 2010 y 2012 de ELPI.

Los resultados permiten cumplir con el objetivo de lograr interpretaciones longitudinales sobre el desarrollo infantil, utilizando datos poblacionales. Este objetivo es posible ya que se puede interpretar el desarrollo infantil desde la trayectoria que es. Todo esto es posible gracias a que el escalamiento vertical puede generar una escala común.

Considerando estos antecedentes, a modo de ejemplo, se interpreta los resultados de un caso presente en ambas rondas de ELPI. El caso folio 102731 cuenta con 22 y 50 meses de edad en las rondas de ELPI 2010 y 2012 respectivamente.

Siguiendo una interpretación en base a norma, el puntaje del caso sería 42 y 44 para los años 2010 y 2012 respectivamente. Cabe consignar que estos puntajes fueron extraídos directamente de las bases de datos disponibles. Corresponden a puntajes estandarizado considerando la muestra de cada ronda. (Centro Micro Datos, s.f.-e) (Centro Micro Datos, s.f.-b)

Tal como se detalló en esta tesis, este tipo de interpretaciones no permite dar cuenta de la trayectoria del desarrollo, ni devela que aspectos son los que la niña o niño es capaz de desenvolver en su ambiente.

Por el contrario, considerando el parámetro de habilidad EAP obtenido gracias al procedimiento de escalamiento vertical, el mismo caso tenía una habilidad de -0,3632 en 2010 y de 1,0513 en 2012. Ambas expresiones están en la misma escala logit.

El hecho de que ambas puntuaciones estén en la misma escala permite develar el crecimiento de la variable a través del tiempo. Se puede saber exactamente la magnitud del cambio entre cada ronda. En este caso es de 1,4145 logit. Esto es un símil a la talla del mismo

caso, en donde la base datos nos informa que al año 2010 media 87 centímetros y en el 2012 media 105 centímetros, con lo cual creció 18 centímetros.

De igual manera, la elección del modelo de Rasch, permite obtener interpretaciones sustantivas de estos resultados por tener en la misma escala tanto la habilidad de las personas como la dificultad de los ítems. Es así como, a partir de los resultados obtenidos, el caso folio 102731 en cuanto a su comunicación expresiva en la ronda 2010 es probable que utilizará 10 palabras, no necesariamente fonéticamente correctas (tales como “aba” para agua o “guau-guau” para perro), para designar a personas u objetos. En tanto en la ronda 2012 el caso es capaz de utilizar correctamente los pronombres “yo”, “tú” y “mí” diariamente.

Al menos dos ventajas se obtienen de este tipo de interpretaciones. En primer lugar, se pueden entender los rezagos en el desarrollo infantil según criterios sustantivos. Con la interpretación bajo normas lo catalogado como rezago en el desarrollo está dado según ciertos puntos de corte en los puntajes. Por ejemplo -2 desviaciones estándar es considerado un rezago. Pero, encontrarse en ciertos puntos de una distribución de puntajes no esclarece que aspectos son los deficitarios en el desarrollo de cada sujeto.

En segundo lugar, se accede a analizar el crecimiento y distribución del desarrollo infantil en distintas etapas del ciclo vital y entre cohortes. Con la interpretación bajo normas la información se limita a la comparación con el grupo de referencia. No se puede hacer un análisis de toda la trayectoria del desarrollo infantil en su conjunto. Además, en el caso específico de los puntajes estandarizados, la distribución queda fijada con una media de 0 y desviación estándar de 1. Por lo tanto, se cierra la posibilidad de observar distintas distribuciones entre diversas etapas del desarrollo o entre cohortes.

Es así como, las interpretaciones longitudinales y sustantivas del constructo de desarrollo infantil son relevantes ya que permiten tomar decisiones informadas y con sustentos claros sobre la primera infancia en Chile. De igual manera, esto abre un abanico de posibilidades en relación al estudio del desarrollo infantil para explicar desventajas, promotores del desarrollo o efectos de situaciones adversas con el objetivo de contribuir a un mejor bienestar de la primera infancia.

Sin embargo, hay que destacar que una dificultad de este trabajo es el cambio de instrumento entre rondas. Se trabajó sobre el supuesto que los instrumentos BDI y BDI-2 Screening conceptualmente miden lo mismo y que corresponden a versiones del mismo

instrumento. Por ende, se tenía la suposición de que existían ítems comunes entre ambos instrumentos.

Finalmente, el número de ítems comunes utilizados fue sólo de 13. A pesar de que esta solución fue estable en comparación a soluciones con mayor número de ítems, no es una situación ideal. Cómo se detalló, los ítems comunes funcionan como anclas para ambas pruebas, proveyendo información para estas dos. Al existir menos ítems anclas, menor es la información común y se vuelve más inestable la estimación en su conjunto.

Consecuentemente, el hallazgo de escasos ítems comunes entre las versiones BDI y BDI-2 Screening utilizados en ELPI, hace que ambos instrumentos sean operacionalmente distintos. Por ende, cualquier comparación de puntajes entre ambos, sin escalamiento, no sería correcto.

Para futuras rondas de ELPI se recomienda no variar los instrumentos de medición de desarrollo infantil. En el caso que se haga necesario un escalamiento vertical, se recomienda tener presente que los ítems comunes deben representar al menos el 20% de algunas de las pruebas a escalar (Kolen & Brennan, 2014)

## **Bibliografía**

- Bedregal, P. (2008). Instrumento de medición del desarrollo en Chile. *Revista Chilena de Pediatría*, 79, 32-36.
- Bedregal, P., & Pardo, M. (2004). *Desarrollo Infantil Temprano y Derechos del niño*. Santiago-Chile: UNICEF.
- Berls, A. T., & McEwen, I. (1999). Battelle developmental inventory. *Physical therapy*, 79(8), 776-783.
- Bliss, S. (2007). Test Reviews: Newborg, J.(2005). Battelle Developmental Inventory– Second Edition. Itasca, IL: Riverside. *Journal of Psychoeducational Assessment*, 25(4), 409-415.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.
- Briggs, D., & Weeks, J. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3-14.
- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, Massachusetts: Harvard university press.
- Centro Micro Datos. (s.f.-a). *Informe Resultados Evaluaciones Primera Ronda Encuesta Longitudinal Primera Infancia*. Santiago.
- Centro Micro Datos. (s.f.-b). *Informe Resultados Evaluaciones Segunda Ronda Encuesta Longitudinal de la Primera Infancia*. Santiago.
- Centro Micro Datos. (s.f.-c). *Manuel de usuario base de datos primera ronda encuesta longitudinal de la primera infancia*.
- Centro Micro Datos. (s.f.-d). *Manual usuario base de datos segunda ronda encuesta longitudinal de la primera infancia*.
- Centro Micro Datos. (s.f.-e). *Encuesta Longitudinal de la Primera Infancia Anexo metodológico instrumentos utilizados para la evaluación*.
- Centro UC Encuestas. (2018). *Reporte Metodológico Encuesta Longitudinal Primera Infancia (III ronda)* . Santiago de Chile.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage learning .
- Elbaum, B., Gattamorta, K., & Penfield, R. (2010). Evaluation of the Battelle developmental inventory, screening test for use in states' child outcomes measurement systems under the Individuals with Disabilities Education Act. *Journal of Early Intervention*, 32(4), 255-273.
- Fernald, L., Kariger, P., Engle, P., & Raikes, A. (2009). *Examining Early Child Development in Low-Income Countries: A Toolkit for the Assessment of Children in the First Five Years of Life*. World Bank.
- Flavell, J. H., Miller, P., & Miller, S. (1993). *Cognitive Development*. Englewood Cliffs: Prentice Hall.

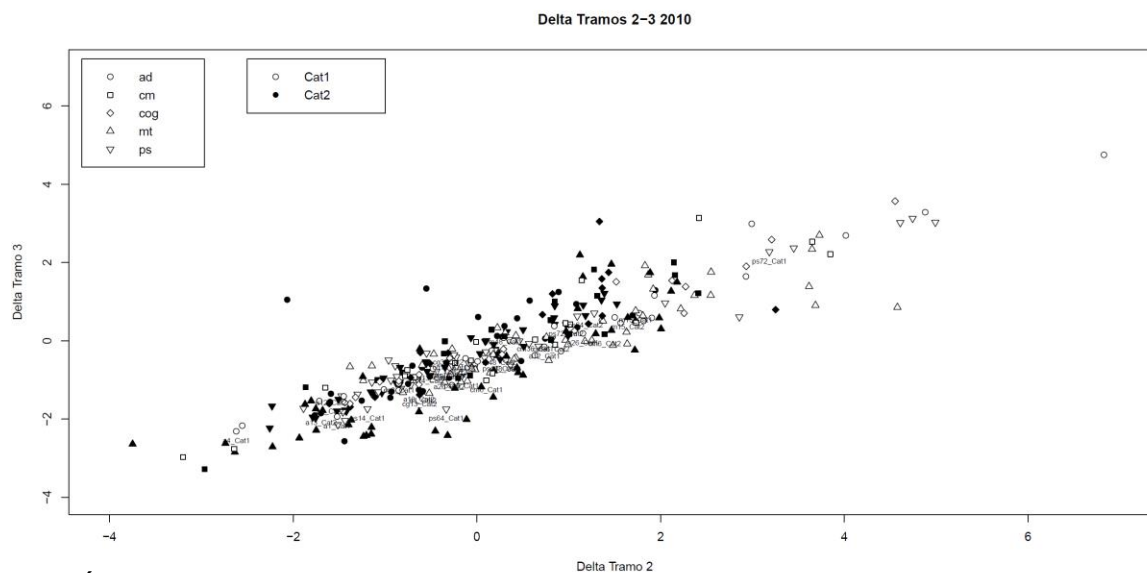
- Gilliam, W., & de Mesquita, P. (2000). The relationship between language and cognitive development and emotional-behavioral problems in financially-disadvantaged preschoolers: A longitudinal investigation. *Early Child Development and Care*, 162(1), 9-24.
- Glascoe, F. P., & Byrne, K. (1993). The usefulness of the Battelle developmental inventory screening test. *Clinical pediatrics*, 32(5), 273-280.
- Harris, D. (2007). Practical issues in vertical scaling. En N. Dorans, M. Pommerich, & P. Holland, *In Linking and aligning scores and scales* (págs. 233-251). New York: Springer.
- Hurt, H., Malmud, E., Betancourt, L., Brodsky, N., & Giannetta, J. (2001). A prospective comparison of developmental outcome of children with in utero cocaine exposure and controls using the Battelle Developmental Inventory. *Journal of Developmental & Behavioral Pediatrics*, 22(1), 27-34.
- Jentzsch, C. (1994). *The predictive validity of the Battelle Developmental Inventory as a measure of adaptive behavior: A 2-3 year, longitudinal comparison with the Scales of Independent Behavior*.
- Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer Science & Business Media.
- Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en educación médica*, 3(9), 40-55.
- Mariscal, S. (2008). El desarrollo del lenguaje oral. En A. Corral, B. Delgado, M. García, M. Giménez, & S. Mariscal, *Psicología del desarrollo desde el nacimiento a la primera infancia* (págs. 209-235). España: McGraw-Hill.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Medeiros, K., Cress, C., & Lambert, M. (2016). Mastery motivation in children with complex communication needs: longitudinal data analysis. *Augmentative and Alternative Communication*, 32(3), 208-218.
- Medwave Estudios Limitada. (2017). *Análisis de instrumentos para la evaluación del desarrollo infantil de niños(as) de 0 a 4 años en Chile*. Santiago de Chile.
- Ministerio de Desarrollo Social. (s.f.). *Libro de códigos evaluaciones*.
- Ministerio de Salud. (2018). *Patrones de crecimiento para la evaluación nutricional de niños, niñas y adolescentes, desde el nacimiento hasta los 19 años de edad*. Obtenido de <http://www.bibliotecaminsal.cl/wp/wp-content/uploads/2018/03/2018.03.16-Patrones-de-crecimiento-para-la-evaluaci%C3%B3n-nutricional-de-ni%C3%B1os-ni%C3%B1as-y-adolescentes-2018.pdf>
- Newborg, J., Stock, J., & Wnek, L. (2011). *Inventario de Desarrollo Battelle*. Madrid: tea ediciones.
- Nip, I., Green, J., & Marx, D. (2011). The co-emergence of cognition, language, and speech motor control in early development: A longitudinal correlation study. *Journal of Communication Disorders*, 44(2), 149-160.

- Observatorio Ministerio Desarrollo Social. (s.f.). *ELPI Observatorio Ministerio Desarrollo Social*.  
Obtenido de Antecedentes ELPI 2010-2012:  
[http://observatorio.ministeriodesarrollosocial.gob.cl/elpi/elpi\\_antecedentes.php](http://observatorio.ministeriodesarrollosocial.gob.cl/elpi/elpi_antecedentes.php)
- Patz, R., & Yao, L. (2007). Vertical Scaling: Statistical Models for Measuring Growth and Achievement. En C. R. Rao, & S. Sinharay, *Handbook of Statistics: Psychometrics* (Vol. 26, págs. 955-975). Amsterdam: Elsevier.
- Piaget, J. (1981). La teoría de Piaget. *Infancia y aprendizaje*, 4(sup2), 13-54.
- Piaget, J., & Inhelder, B. (2008). *Psicología del Niño*. Buenos Aires: Morata.
- Raikes, H. A. (2017). *Measuring of child development and learning*.
- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*.
- Rost, J. (2012). The Growing Family of Rasch Models. En A. Boomsma, M. van Duijn, & T. Snijders, *Essays on Item Response Theory* (págs. 25-42). Springer.
- Snyder, P., Lawson, S., Thompson, B., Stricklin, S., & Sexton, D. (1993). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. *Topics in Early Childhood Special Education*, 13(2), 216-232.
- Sroufe, A. (1997). *Emotional development: The organization of emotional life in the early years*. Cambridge University Press.
- Thorndike, R., & Thorndike, C. (2010). *Measurement and evaluation in psychology*. Boston, MA: Pearson.
- Torres Irribarra, D. (2018). Rasch Model. En B. Frey, *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (Vol. 3, págs. 1370-1373). Sage.
- Vigotski, L. (1979). *El desarrollo de los procesos psicológicos superiores*. Barcelona: Crítica.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.
- Young, M., & Tong, Y. (2006). Vertical Scales. En S. Lane, M. Raymond, & T. Haladyna, *Handbook of Test Development* (págs. 469-485). Routledge.

## Anexos

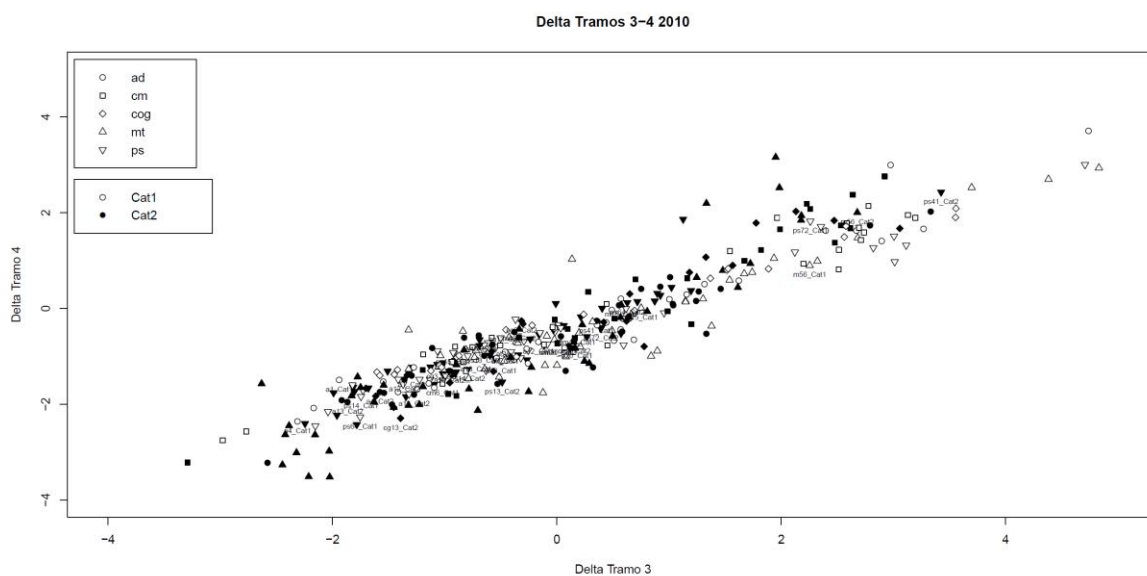
### Anexo N°1

#### Dificultades tramos 2 y 3 BDI año 2010



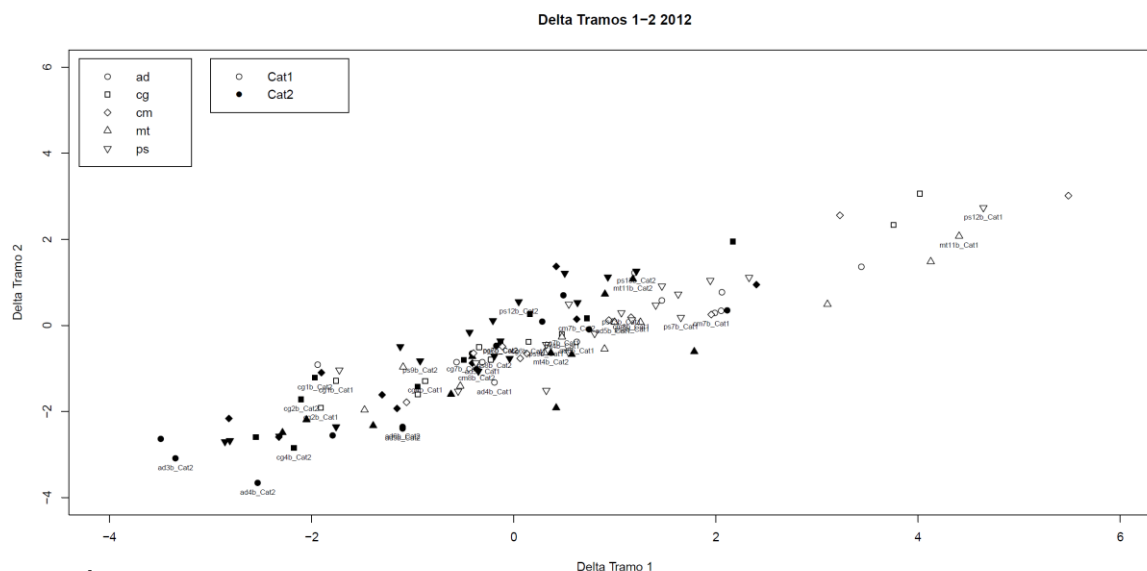
*Nota.* Ítem común PS41 no incluido en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

#### Dificultades tramos 3 y 4 BDI año 2010



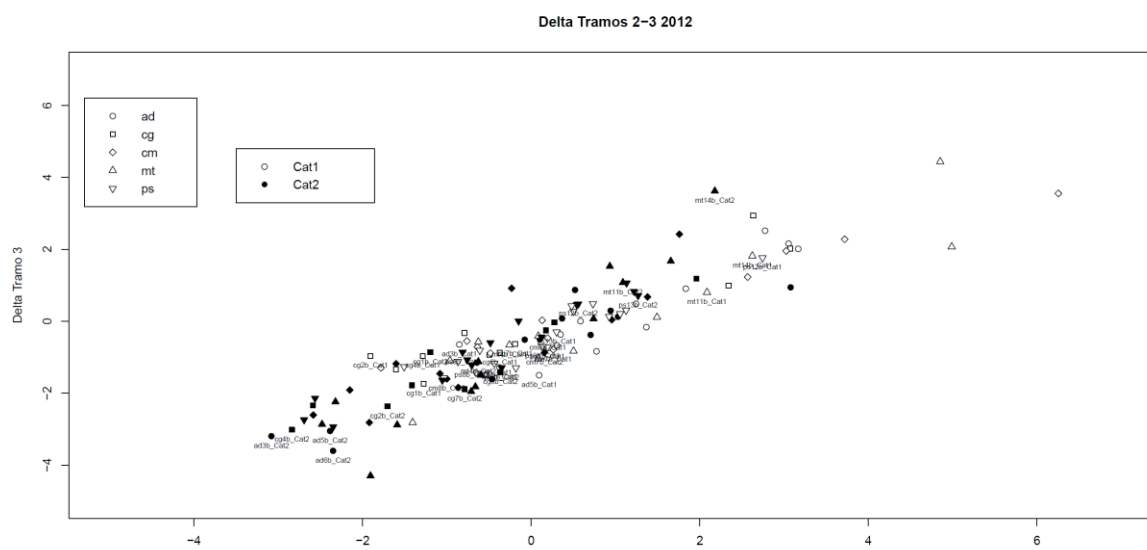
## Anexo N°2

### Dificultades tramos 1 y 2 BDI-2 Screening año 2012



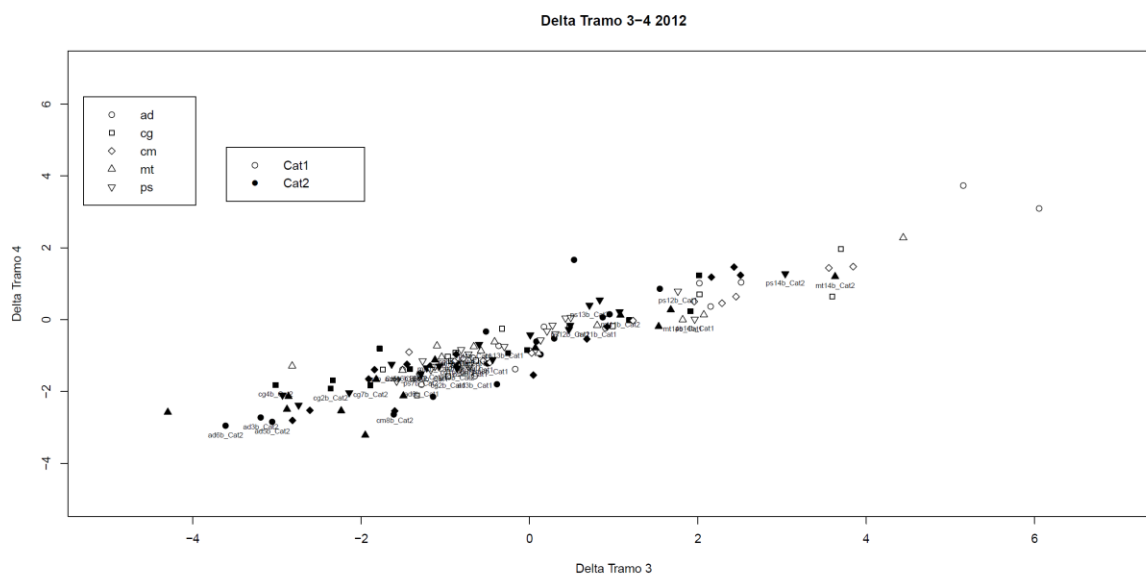
*Nota.* Ítems comunes PS14 y MT14 no incluidos en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

### Dificultades tramos 2 y 3 BDI-2 Screening año 2012



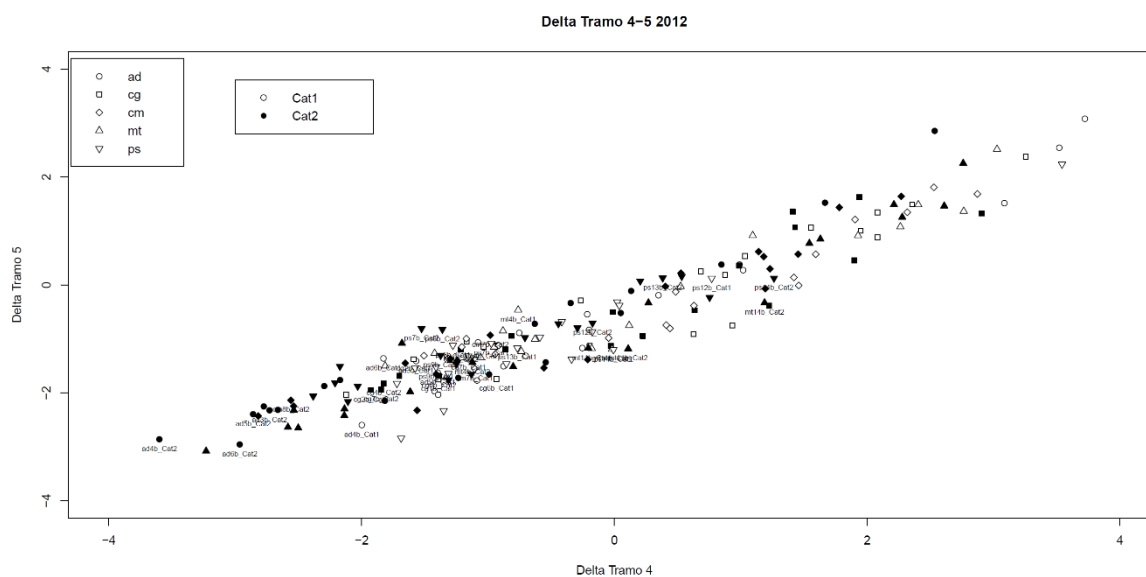
*Nota.* Ítems comunes AD4 y PS14 no incluidos en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

*Dificultades tramos 3 y 4 BDI-2 Screening año 2012*

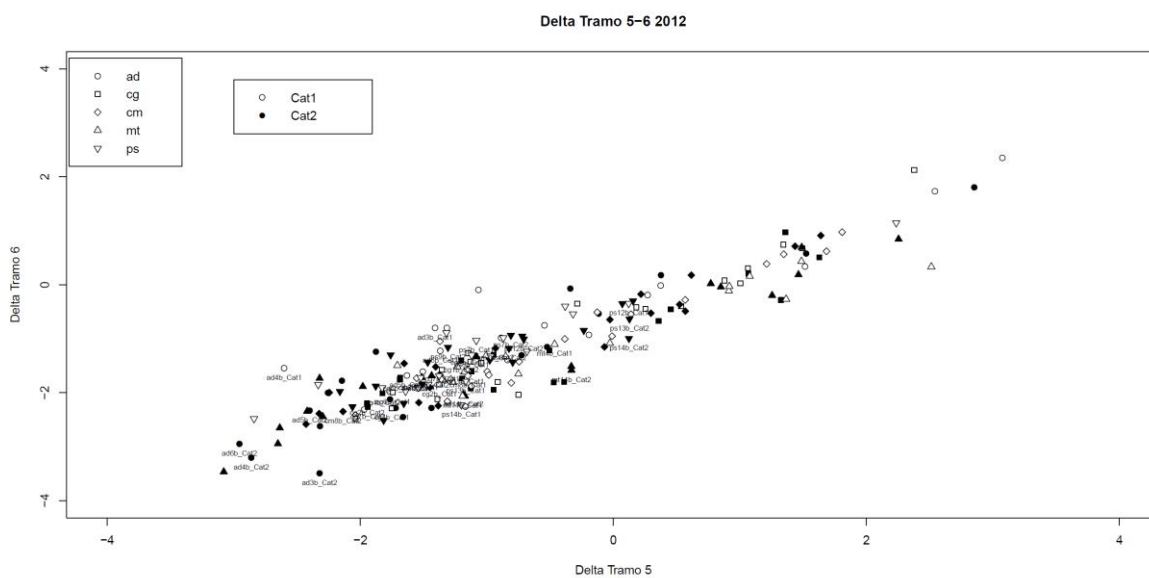


*Nota.* Ítem común AD4 no incluido en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

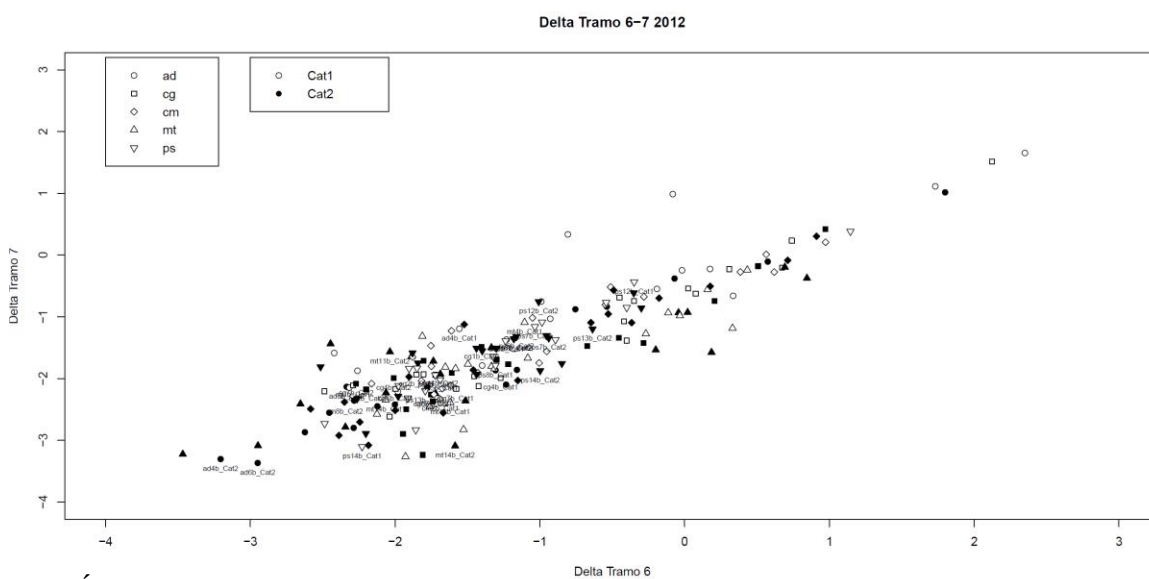
*Dificultades tramos 4 y 5 BDI-2 Screening año 2012*



### Dificultades tramos 5 y 6 BDI-2 Screening año 2012

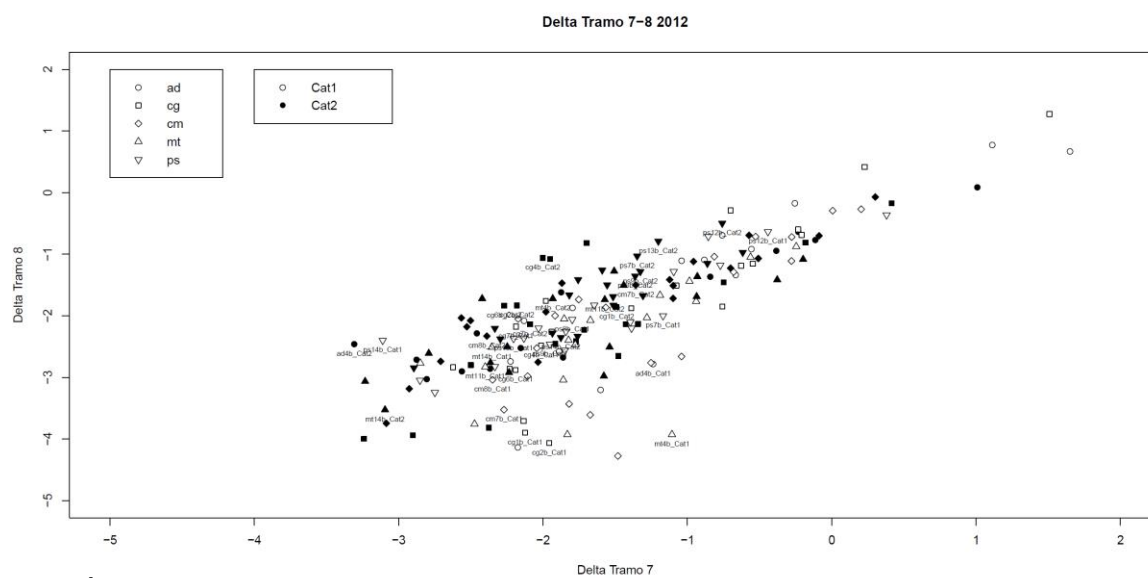


### Dificultades tramos 6 y 7 BDI-2 Screening año 2012



*Nota.* Ítem común AD3 no incluido en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

*Dificultades tramos 7 y 8 BDI-2 Screening año 2012*



*Nota.* Ítems comunes AD3, AD5, AD6 no incluidos en el gráfico por tener una frecuencia inferior a 10 casos en al menos dos categorías de respuesta.

**Anexo N°3**

| Nombre item | año  | Escalamiento 1 | Escalamiento final | Escalamiento 2 | Escalamiento 3 |
|-------------|------|----------------|--------------------|----------------|----------------|
| ad4b        | 2012 | e1             | e1                 | e1             | e1             |
| a13         | 2010 | e1             | e1                 | e1             | e1             |
| ad6b        | 2012 | e2             | e2                 | e2             | e2             |
| a18         | 2010 | e2             | e2                 | e2             | e2             |
| ps8b        | 2012 | e3             | e3                 | e3             | e3             |
| ps45        | 2010 | e3             | e3                 | e3             | e3             |
| ps9b        | 2012 | e4             | e4                 | e4             | e4             |
| ps14        | 2010 | e4             | e4                 | e4             | e4             |
| ps12b       | 2012 | e5             | e5                 | e5             | e5             |
| ps72        | 2010 | e5             | e5                 | e5             | e5             |
| ps13b       | 2012 | e6             | e6                 | e6             | e6             |
| ps64        | 2010 | e6             | e6                 | e6             | e6             |
| cm7b        | 2012 | e7             | e7                 | e7             | e7             |
| cm6         | 2010 | e7             | e7                 | e7             | e7             |
| cm8b        | 2012 | e8             | e8                 | e8             | e8             |
| cm36        | 2010 | e8             | e8                 | e8             | e8             |
| mt4b        | 2012 | e9             | e9                 | e9             | e9             |
| m48         | 2010 | e9             | e9                 | e9             | e9             |
| mt11b       | 2012 | e10            | e10                | e10            | e10            |
| m15         | 2010 | e10            | e10                | e10            | e10            |
| mt14b       | 2012 | e11            |                    | e11            | e11            |
| m56         | 2010 | e11            |                    | e11            | e11            |
| cg2b        | 2012 | e12            | e12                | e12            | e12            |
| a1          | 2010 | e12            | e12                | e12            | e12            |
| cg6b        | 2012 | e13            | e13                | e13            | e13            |
| cg13        | 2010 | e13            | e13                | e13            | e13            |
| cg7b        | 2012 | e14            | e14                | e14            | e14            |
| a7          | 2010 | e14            | e14                | e14            | e14            |
| ad3b        | 2012 |                |                    | e15            | e15            |
| a12         | 2010 |                |                    | e15            | e15            |
| ad5b        | 2012 |                |                    | e16            | e16            |
| a26         | 2010 |                |                    | e16            | e16            |
| ps7b        | 2012 |                |                    | e17            | e17            |
| ps13        | 2010 |                |                    | e17            | e17            |
| ps14b       | 2012 |                |                    | e18            | e18            |
| ps41        | 2010 |                |                    | e18            | e18            |
| cg1b        | 2012 |                |                    | e19            | e19            |
| a4          | 2010 |                |                    | e19            | e19            |

|       |      |  |  |     |     |
|-------|------|--|--|-----|-----|
| cg4b  | 2012 |  |  | e20 | e20 |
| cg3   | 2010 |  |  | e20 | e20 |
| ad2b  | 2012 |  |  |     | e21 |
| a14   | 2010 |  |  |     | e21 |
| cm3b  | 2012 |  |  |     | e22 |
| cm4   | 2010 |  |  |     | e22 |
| cm4b  | 2012 |  |  |     | e23 |
| cm30  | 2010 |  |  |     | e23 |
| cm6b  | 2012 |  |  |     | e24 |
| cm35  | 2010 |  |  |     | e24 |
| cm10b | 2012 |  |  |     | e25 |
| cm38  | 2010 |  |  |     | e25 |
| mt5b  | 2012 |  |  |     | e26 |
| m37   | 2010 |  |  |     | e26 |
| mt9b  | 2012 |  |  |     | e27 |
| m40   | 2010 |  |  |     | e27 |

