Actualidades en Psicología
ISSN: 0258-6444
actualidades.psicologia@ucr.ac.cr
Instituto de Investigaciones Psicológicas
Costa Rica

Montenegro Montenegro, Esteban; Oh, Youngha; Chesnut, Steven
No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos
Actualidades en Psicología, vol. 29, núm. 119, 2015, pp. 29-42
Instituto de Investigaciones Psicológicas
Jan sosé, Costa Rica

Actualidades en Psicología, 29(119), 2015, 29-42 ISSN 2215-3535
http://revistas.ucr.ac.cr/index.php/actualidades DOI: http://dx.doi.org/10.15517/ap.v29i119.18812

No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos

Do not Be Afraid of Missing Data: Modern Approaches to Handle Missing Information

Esteban Montenegro Montenegro[1]
Youngha Oh[2]
Texas Tech University, United States
Steven Chesnut[3]
University of Southern Mississippi, United States

Resumen. La mayoría de los datos en ciencias sociales y educación presentan valores perdidos debido al abandono del estudio o la ausencia de respuesta. Los métodos para el manejo de datos perdidos han mejorado dramáticamente en los últimos años, y los programas computacionales ofrecen en la actualidad una variedad de opciones sofisticadas. A pesar de la amplia disponibilidad de métodos considerablemente justificados, muchos investigadores e investigadoras siguen confiando en técnicas viejas de imputación que pueden crear análisis sesgados. Este artículo presenta una introducción conceptual a los patrones de datos perdidos. Seguidamente, se introduce el manejo de datos perdidos y el análisis de los mismos con base en los mecanismos modernos del método de máxima verosimilitud con información completa (FIML, siglas en inglés) y la imputación múltiple (IM). Asimismo, se incluye una introducción a los diseños de datos perdidos así como nuevas herramientas computacionales tales como la función Quark y el paquete semTools. Se espera que este artículo incentive el uso de métodos modernos para el análisis de los datos perdidos.

Palabras clave. datos perdidos, máxima verosimilitud con información completa, imputación múltiple, diseños de datos perdidos, psicometría.

Abstract. Most of the social and educational data have missing observations due to either attrition or nonresponse. Missing data methodology has improved dramatically in recent years, and popular computer programs as well as software now offer a variety of sophisticated options. Despite the widespread availability of theoretically justified methods, many researchers still rely on old imputation techniques that can create biased analysis. This article provides conceptual introductions to the patterns of missing data. In line with that, this article introduces how to handle and analyze the missing information based on modern mechanisms of full-information maximum likelihood (FIML) and multiple imputation (MI). An introduction about planned missing designs is also included and new computational tools like Quark function, and semTools package are also mentioned. The authors hope that this paper encourages researchers to implement modern methods for analyzing missing data.

Keywords. missing data, maximum likelihood estimation, full-information maximum likelihood, multiple imputation, planned missingness, psychometrics.

[1]Esteban Montenegro-Montenegro. Institute for Measurement, Methodology, Analysis and Policy, Texas Tech University. Dirección postal: Texas Tech University - National Wind Institute 1009 Canton Ave. Room Number 211 Lubbock, TX 79409, United States. Email: esteban.montenegro@ttu.edu
[2]Youngha Oh. Institute for Measurement, Methodology, Analysis and Policy, Texas Tech University, United States. Email: youngha.oh@ttu.edu
[3]Steven Chesnut. University of Southern Mississippi, United States. Email: steven.chesnut@usm.edu

Introducción

La presencia de datos perdidos ha sido siempre considerada un problema en el campo de la medición, especialmente en psicología y educación, debido a que disminuye el poder estadístico. Asimismo, el enfoque clásico de manejo de datos perdidos plantea una serie de medidas poco efectivas como la eliminación de casos cuyos ítems fueron no respondidos o la sustitución por la media (Enders, 2010; Baraldi & Enders, 2010).

En la actualidad se puede contar con enfoques para el manejo de datos perdidos más eficientes y fáciles de implementar gracias a los avances en computación (Enders, 2010; Graham, 2012), permitiendo así recuperar los valores perdidos y restablecer el poder estadístico. Debido a esto, el principal problema ya no es la presencia de valores perdidos, el verdadero problema es como lidiamos con los datos perdidos (Little, Jorgensen, Lang, & Moore, 2014).

Es frecuente en los tiempos actuales, encontrar investigadores e investigadoras que consideran los métodos modernos de manejo de datos perdidos como "un engaño" o incluso "inmoral" tal como lo relata Little (2013). Esto demuestra el escaso conocimiento acerca del alcance de estas técnicas modernas y la existencia de programas de computadora; incluso gratuitos, para el manejo de valores perdidos.

Asimismo, los datos perdidos no son más un problema que debe ser prevenido, ya que es posible utilizar diseños que contemplen perder datos de manera deliberada en aras de economizar tiempo y dinero sin arriesgar en gran medida el poder estadístico (Graham, 2006; Rhemtulla & Little, 2012). Así, es factible abarcar muestras numerosas con instrumentos extensos sin agotar las energías y recursos cognitivos de los y las participantes (Little et al., 2014).

El propósito de la presente revisión es difundir los hallazgos más actuales en el manejo de datos perdidos, asimismo, ante la escasa literatura al respecto en español, se busca crear un texto sencillo y de fácil comprensión acerca de los supuestos que subyacen los procedimientos más utilizados para la sustitución y manejo de datos perdidos. Otro de los objetivos es introducir el diseño de estudios con datos perdidos previamente planificados y su uso en la investigación en psicología y educación, especialmente para la indagación de las propiedades psicométricas de los ítems utilizando muestras numerosas. Por último, se pretende introducir nuevas herramientas computacionales que pueden ser de ayuda para el manejo de los datos perdidos o para su previa planificación.

Es importante resaltar que el tema será tratado en el contexto del modelado de ecuaciones estructurales debido a su amplio uso en ciencias sociales y en especial para la medición en psicología y educación (Little, 2013; Kline, 2010)

Mecanismos de datos perdidos

La mayoría de datos en ciencias sociales y educación presentan datos perdidos debido a la ausencia de respuesta por parte de los participantes o debido a que los y las participantes abandonan el estudio. De acuerdo a Rubin (1976), los patrones de datos perdidos pueden clasificarse; según la relación entre los datos perdidos y los datos, en datos perdidos completamente al azar (missing completely at random [MCAR]), datos perdidos al azar (missing at random [MAR]) y datos perdidos no aleatorios (missing not at random [MNAR]). Los datos perdidos pueden considerarse completamente perdidos al azar cuando no están relacionados con ninguna variable presente o no en los datos (Little, et al., 2014). Supongamos que solo la variable Y contiene valores perdidos, y tenemos otro grupo de variables representadas por el vector *X*. Los datos podrían considerarse perdidos completamente al azar si la probabilidad de valores perdidos en *Y* no depende de *X* o *Y* en sí misma; y si no depende de otra variable no presente en los datos. (Rubin, 1976). Para representar esto formalmente, consideremos R como la "respuesta" que puede adoptar un valor de "1" si Y tiene datos perdidos y "0" si Y tiene datos completos. Así, datos perdidos completamente al azar significaría:

$$\Pr(R = 1 \mid X, Y) = \Pr(R = 1) \qquad (1)$$

Un ejemplo más simple, siguiendo la anterior afirmación, sería suponer que Y es una medida de delincuencia y X sería años de escolaridad, en este caso MCAR se cumpliría si la probabilidad de encontrar

datos perdidos en la medida de delincuencia no está relacionada con años de escolaridad o con la variable en sí misma (u otra variable no presente en los datos). Esto es de especial relevancia, ya que la mayoría de técnicas tradicionales para el manejo de datos perdidos requieren el cumplimiento de este supuesto. Si bien es difícil cumplir el supuesto de MCAR en contextos de escaso control, hay ciertas situaciones en las que MCAR es posible de sostener. El mecanismo de datos perdidos MCAR es considerado el escenario ideal de datos perdidos debido que es un proceso totalmente aleatorio donde los valores perdidos son totalmente arbitrarios y no habría sesgo alguno (Little, et al., 2014; Little, 2013; Rhemtulla & Little, 2012).

Otro de los mecanismos de datos perdidos; MAR supone que los datos perdidos existen por una razón predecible, de esta forma los datos perdidos son causados por un efecto aleatorio fácilmente estimable. Siguiendo el ejemplo anterior, supongamos que tenemos una variable *Y* la cual contiene valores perdidos, y a su vez tenemos una variable *X* que no contiene valor perdido alguno. En este caso podríamos afirmar que los datos perdidos en *Y* fueron perdidos aleatoriamente si la probabilidad de que *Y* tenga valores perdidos no depende de *Y*, una vez que controlamos el efecto de *X*. De manera formal, podemos expresar esta afirmación como:

$$\Pr(R = 1 \mid X, Y) = \Pr(R = 1 \mid X) \qquad (2)$$

Así, el supuesto de MAR permitiría que los datos perdidos en Y dependan de otra variable, pero los datos perdidos en Y no pueden depender de la variable en sí misma. Asimismo, continuando con el ejemplo anterior, si Y fuera una medida de delincuencia y X fuera años de escolaridad, el criterio de MAR se cumpliría si los datos perdidos de la variable sobre delincuencia dependiera en años de escolaridad. En principio, bajo el supuesto de MAR, los valores perdidos dependerían de variables observadas, y no dependería en la noción intuitiva de efecto aleatorio (Little et al., 2014). Otro ejemplo posible, para entender el supuesto MAR, es la tendencia de los varones a negarse a responder estudios acerca de depresión, pero esta tendencia no tiene relación con su nivel depresión, sino más bien; con otras variables de socialización (Little et al., 2014).

Finalmente, el tercer mecanismo de datos perdidos sería los datos perdidos no aleatorios (MNAR). Cuando una variable cumple el supuesto de MNAR, la razón para la existencia de datos perdidos subyace en la variable en sí misma, lo cual significa que este mecanismo ocurre cuando los datos perdidos, en una determinada variable, ocurren debido a los niveles de los sujetos en esa variable (Little et al., 2014). En este caso, no habría mayor información disponible para estimarlos, si lo que se desea es recuperar dichos valores. Un buen ejemplo sería un estudio acerca de consumo de tabaco en adolescentes quienes no suelen reportar cuantos cigarros consumen. En este ejemplo, el hecho de que fumar sea ilegal si no se cumple la mayoría de edad aceptada, hace que los y las participantes teman algún tipo de represalia por parte de sus padres o problemas legales. De esta forma, la pregunta en si misma sería la causa de los datos perdidos y esto haría difícil la labor de recuperar la información ya que la causa de los valores perdidos no podría ser usada para corregir el sesgo de los parámetros estimados, como si podría ser posible si se cumple el supuesto de MCAR o MAR.

Tratamiento de datos perdidos

Tal como se sugirió anteriormente, existen diferentes supuestos acerca de la relación de los valores perdidos en un estudio y los datos recolectados. El cumplimiento de los supuestos MCAR o MAR son los requisitos necesarios para realizar la recuperación de los datos perdidos. Asimismo, existen técnicas de sustitución de datos de más larga data tales como sustitución por la media o regresión, que requerirían el cumplimiento del supuesto de MCAR, lo cual es una condición difícil de cumplir al menos que los datos perdidos hayan sido incluidos como parte del diseño (tema abordado más adelante). Si se procede con un enfoque clásico sin tener datos perdidos con un patrón MCAR, estimaciones tales como correlaciones, diferencias de medias, etc., serán muy grandes o se verían atenuados, asimismo los errores estándar de las pruebas de significancia serían menores al introducir valores artificiales tales como la media de la variable

o la media del grupo (Enders 2010; Graham, 2009, 2012; Little et al., 2014; Little, 2013).

Para evitar este tipo de errores y aprovechar las bondades de los datos perdidos al azar, existen diferentes enfoques modernos para el tratamiento de los valores perdidos, uno de ellos es el método de máxima verosimilitud con información completa (FIML, siglas en inglés) e imputación múltiple (IM) (Arbuckle, 1996; Enders & Bandalos, 2001).

Imputación múltiple

La imputación[1] múltiple consiste en hacer copia de la base de datos original y reemplazar los valores perdidos con estimaciones probables de los valores que hubieran existido en las celdas vacías, si fueran valores observados (Rubin, 1987). La aplicación de esta técnica requiere de tres pasos: imputación, análisis y pooling (integración de los valores correspondientes a todas las copias generadas) (Allison, 2001). El primer paso de la imputación múltiple es el más complicado del proceso y además existen varias formas para realizarlo. Una de las estrategias más populares es la sustitución de valores perdidos utilizando regresión de imputación. Supongamos que una base datos tiene un conjunto de variables *X, Y* y *Z*. También, asumamos que *X* y *Y* no poseen ningún valor perdido, sin embargo *Z* presenta un 20% de casos perdidos. Para sustituir los datos perdidos en la variable *Z*, se efectúa una regresión de las variables *X* y *Y* en *Z*, la siguiente ecuación representaría este primero paso:

$$\hat{Z} = b_0 + b_1 X + b_2 Y \qquad (3)$$

La regresión convencional de imputación simplemente introduciría valores para *X* y *Y* para los casos con datos perdidos y estimaría valores para la variable *Z*. No obstante, estos valores introducidos tendrían una varianza muy pequeña, lo cual causaría sesgo en otros parámetros estimados. Para corregir este problema, se puede utilizar la siguiente ecuación:

$$\hat{Z} = b_0 + b_1 X + b_2 Y + sE \qquad (4)$$

Donde E es un valor seleccionado aleatoriamente de una distribución normal estándar (con media cero y desviación estándar 1) y s es la desviación estándar estimada del error en la regresión (en este caso la media cuadrática del error). Al añadir esta selección aleatoria, la varianza de los valores sustituidos incrementa y previene los sesgos causados usualmente por el método convencional de imputación (Allison, 2001).

El sesgo en la estimación de los parámetros no es el único problema por solucionar, de ser ese el caso, una sola copia de la base de datos sería suficiente. Sin embargo, una sola base de datos con valores sustituidos sólo estimaría los errores estándar reducidos, además los parámetros estimados no serían estadísticamente eficientes debido a que la variación aleatoria introducida agrega variabilidad muestral. Para hacer frente a esta dificultad, se producen varios conjuntos de bases de datos a partir de la base de datos original. Cada nueva base de datos contiene diferentes valores imputados a partir de la selección aleatoria producida en *E*. Así, el modelo deseado es estimado en cada base de datos, y los parámetros estimados son promediados a través de las múltiples bases de datos. Este proceso conlleva una estimación de parámetros más estable con mayor eficiencia (Allison, 2001).

Con esta estrategia, también se resuelve el problema de los errores estándar al calcular la varianza de cada parámetro entre las diferentes bases de datos generadas. Esta varianza "entre" las bases de datos es la estimación de la variabilidad producida por el proceso de imputación. La varianza "intra" base de datos sería la media de los errores estándar al cuadrado extraída del análisis en cada base de datos. Posteriormente, se obtiene el error estándar ajustado por imputación al estimar la raíz cuadrada de la suma de las varianzas entre e intra. De esta forma, la fórmula para la estimación del error estándar de la media de parámetro de interés (a) sería (Rubin 1987):

[1]La palabra "imputación" se utilizará como sinónimo de sustitución de valores perdidos debido a su frecuente uso con ese sentido en el español.

$$EE(\bar{a}) = \sqrt{\frac{1}{M}\sum_{k=1}^{M} S_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{k=1}^{M}(a_k - \bar{a})^2} \quad (5)$$

En la anterior formula, *M* es el número de bases de datos generados, s_k es el error estándar en k^ésima^ base de datos, a_k sería la estimación del parámetro en la k^ésima^ base de datos, y $\bar{a}$ es la media de los parámetros estimados y el factor (1+1/M) corrige la ecuación debido a que el número de bases de datos es finito.

Respecto a la cantidad de bases de datos necesarias, con una cantidad moderada de datos perdidos, cinco bases de datos son suficientes para producir parámetros eficientes. No obstante, se necesitan más de cinco bases de datos para generar buenas estimaciones de los errores estándar y otros estadísticos asociados, especialmente cuando la fracción de datos perdidos es grande (Allison, 2001; Enders, 2010).

Método de máxima verosimilitud con información completa

El método de máxima verosimilitud con información completa (de ahora en adelante FMIL por sus siglas en inglés) es otra estrategia moderna para el tratamiento de datos perdidos en ciencias sociales. Si aplicamos un cuestionario, pero algunos de los ítems presentan valores perdidos, igual podríamos estimar un modelo que nos permita obtener conclusiones acertadas acerca de la totalidad de la muestra. Este método permite ejecutar el análisis deseado al utilizar los valores observados para ayudar a recuperar la información perdida debido a los datos perdidos (Little et al., 2014). La estimación FMIL estima el logaritmo de verosimilitud para cada individuo basado en las variables presentes en el modelo. Al utilizar solo la información conocida a partir de los datos observados, FMIL puede inferir como debería lucir el modelo sin necesidad de conocer cuál podría ser el valor perdido (Little et al., 2014). De esta forma, FMIL puede ser utilizado con bases de datos con datos perdidos y producir estimaciones que describen correctamente toda la muestra.

Este método ha demostrado en numerosos estudios de simulación que al incluir las variables relacionadas con los valores perdidos, permiten la estimación de parámetros menos sesgados y errores estándar más exactos y las diferencias en comparación a IM son menores (Enders, 2010; Schafer & Graham, 2002)

Por otra parte, sin importar el mecanismo de datos perdidos es importante utilizar el mejor método posible disponible, por ejemplo cuando los datos perdidos siguen un patrón MAR, enfoques ad hoc que intentan solucionar la existencia de valores perdidos sin tomar en cuenta la estructura de los datos (listwise, pairwise, sustitución por la media, imputación condicional media, etc.) deben ser evitados en todas las situaciones.

En el contexto del modelado de ecuaciones estructurales (SEM, por sus siglas en inglés), FIML ha demostrado producir parámetros no sesgados y una estimación menos sesgada de los errores estándar bajo el supuesto de datos perdidos MCAR y MAR (Graham 2003), haciendo la salvedad de que bajo el supuesto de MAR debe incluirse la o las variables relacionadas con la presencia de valores perdidos ya sea en el modelo o como variables auxiliares. En este caso el proceso estima la función de probabilidad para cada individuo basado en las variables que han sido incluidas así toda la información disponible es utilizada. De acuerdo a Graham (2003) un ejemplo podría ser una base datos con 389 casos pero algunas variables solo tienen información para 320 casos. La información acerca del ajuste del modelo es obtenida al sumar la función de ajuste para cada caso, y así la información de ajuste obtenida se basaría en la totalidad de los 389 casos.

En el marco del modelado de ecuaciones estructurales, FMIL no estima solo un valor de chi cuadrado (χ^2), sino más bien, estimaría dos valores distintos de χ^2 correspondientes a dos modelos distintos. El primero correspondería al modelo nulo o sin restricción (modelo H_0), en este modelo las variables no están correlacionadas, mientras el segundo modelo, sería el modelo especificado por el o la investigadora (modelo H_1). La diferencia en el logaritmo de verosimilitud (log-likelihood) de los dos modelos es utilizada para estimar el χ^2 del modelo (Graham, 2003).

Aunado a lo anterior, existe bastante evidencia de que utilizar FMIL o incluso IM es una óptima solución cuando los datos perdidos no cumplen el supuesto de MAR y se incluyen variables auxiliares en el modelo que darían cuenta de los valores perdidos (Enders, 2010; Graham, 2003). La inclusión de variables auxiliares tiene un mayor impacto cuando su relación con los valores perdidos es alta ($r < 0.4$) y cuando la proporción de valores perdidos es elevada (25%) (Collins, Schafer, & Cam, 2001; Graham, 2003).

Existen dos formas de modelar las variables auxiliares, la primera puede ser incluirlas como variables dependientes o la segunda, como variables correlacionadas. Ambos enfoques son equivalentes en efectividad reduciendo el sesgo en los parámetros, no obstante al incluir las variables auxiliares como correlaciones, se obtiene un mejor corrección en los sesgos de ajuste en el modelo (Graham, 2003).

Finalmente, ambos métodos IM y FMIL tienden a generar resultados muy similares. La decisión de cuál método utilizar depende en gran medida de la complejidad del modelo que se desea trazar o la pregunta que se desea responder una vez los datos han sido recuperados. Una ventaja del manejo de datos perdidos utilizando FMIL es la estimación de la interacción de variables (por ejemplo hipótesis de moderación) ya que al ser un alcance basado en las variables incluidas en el modelo, la interacción sería una variable más en el modelo. En cambio, si se desea utilizar IM es necesario especificar un modelo de imputación que conserve la interacción, por ejemplo centrar la variable por su media (Enders, 2010), lo cual hace de FMIL un métodos más sencillo de utilizar.

Por otra parte, FMIL es limitado en la cantidad de variables auxiliares que pueden ser utilizadas, mientras IM es más flexible y permite mayor uso de variables auxiliares que pueden dar cuenta de los valores perdidos. No obstante, si se desea trazar modelos que incluyan puntajes totales de escalas, como es común en psicología, el método más flexible sería la imputación múltiple, ya que FMIL no permitiría estimar una nueva variable a partir de la suma de otras variables (Enders, 2010; Gottschall, West, & Enders, 2012). Sin embargo, esto no sería un problema a tomar en cuenta si se emplea un modelo de ecuaciones estructurales donde los ítems formarían parte de un constructo latente, evitando el uso de la sumatoria de ítems.

Inclusión de variables auxiliares

Tanto en el enfoque de IM y de FMIL es siempre deseable incorporar variables auxiliares en el proceso de imputación o en el modelo. Las variables auxiliares son aquellas que no se planea incluir en el modelo pero están, al menos, moderadamente correlacionadas con las variables que poseen valores perdidos. Al incluir este tipo de variables en el modelo de imputación, se puede reducir la incertidumbre y la variabilidad de los valores imputados. Esto a su vez, puede reducir los errores estándar de los parámetros estimados en el modelo final (Allison, 2001).

Para exponer los beneficios de las variables auxiliares, un ejemplo será de utilidad. Asumamos que tenemos una medida W de ingreso anual y tomemos el vector X como un conjunto de variables observadas que serán incluidas en el modelo final en conjunto con W. Asimismo, supongamos que 30% de los casos presentan valores perdidos en la variable W (ingreso anual), además asumamos que tenemos evidencia para sospechar que las personas con ingreso económico alto son aquellas que no respondieron acerca de su ingreso anual (Allison, 2001). Asumiendo que *R* es la respuesta en *W*, este planteamiento se puede expresar de manera formal:

$$\Pr(R=1|X,W)=f(X,W) \quad (6)$$

Esto sería, la probabilidad de encontrar valores perdidos en *W* depende de *X* y *W* en sí misma, lo cual representa un claro incumplimiento del supuesto MAR. Ante esto, supongamos que podemos tener acceso a otro conjunto de variables *Z* que están relacionadas en conjunto con *W*. Este nuevo vector *Z* puede incluir variables tales como coeficiente intelectual, sexo, prestigio laboral, etc. (Allison, 2001). Al introducir estas variables, se espera que la dependencia de la probabilidad de datos perdidos en *W* desaparezca, tal que:

$$\Pr(R=1|X,W,Z)=f(X,Z) \quad (7)$$

En resumen, para reducir el sesgo de estimación y los errores estándar, la inclusión de variables auxiliares es siempre una estrategia recomendada en modelos de imputación múltiple y la implementación de FMIL (Graham, 2003; Allison 2001).

Componentes principales como variables auxiliares

Tal como se mencionó anteriormente, el uso de variables auxiliares conlleva varias ventajas para el proceso de imputación de datos y estimación de los modelos. Estas variables auxiliares no forman parte del modelo teórico pero incrementan la probabilidad de recuperar la información faltante e incrementan el acercamiento del modelo de imputación al criterio de MAR. De esta forma, las variables mejorarían la exactitud de las estimaciones del modelo (Collins et al., 2001; Graham, 2003).

De acuerdo a Howard, Rhemtulla y Little (2015) las variables auxiliares son incluidas de manera diferente en el proceso de MI y FMIL pero producen resultados similares. La imputación múltiple introduce las variables auxiliares en un primero paso donde las imputaciones son generadas con base en el modelo de imputación y en un segundo momento los datos son analizados, sin embargo en el modelo FMIL, solo hay un modelo, así que las variables que predicen los datos perdidos deben ser incorporados en el modelo como variables auxiliares.

Asimismo, existen dos estrategias para incluir variables auxiliares: la primera es la estrategia restrictiva mientras la segunda se considera una estrategia inclusiva. La primera consiste en añadir solo un número reducido de variables auxiliares mientras la estrategia inclusiva conlleva añadir tantas variables auxiliares como sea posible. Para someter a prueba las bondades de ambas estrategias; Collins et al. (2001) efectuaron un estudio de simulación donde se demostró que la estrategia inclusiva ayuda a obtener resultados más eficientes debido a que hay menor probabilidad de omitir accidentalmente una causa importante de datos perdidos, es decir; hay mayor certeza de cumplir el criterio de MAR, además hay también mayor posibilidad de reducir los sesgos de estimación. Sin embargo, no existe hasta el momento, alguna guía para lidiar cantidades grandes de variables auxiliares siguiendo la estrategia inclusiva (Howard, Rhemtulla & Little, 2015).

Para resolver este problema existe una nueva estrategia que propone reducir las variables auxiliares utilizando análisis de componentes principales (ACP) para reducir el número de variables auxiliares. Para ello, se representaría las variables auxiliares como una fuente combinada de información predictiva en lugar de un conjunto de variables individuales (Howard, Rhemtulla & Little, 2015).

La idea principal detrás de ACP es encontrar, a través de una descomposición por valores propios (eigenvalues), un conjunto k de componentes principales ($c1, c2, \ldots, ck$) que contengan la mayor cantidad de varianza posible extraída del grupo original de variables p ($v1, v2, \ldots, vp$) donde $k < p$ sería la combinación lineal de p variables que resultan en una variable con la varianza máxima, y donde $c2$ es la combinación lineal que resulta en la máxima varianza que es ortogonal a $c1$, y así hasta ck componentes. Para elegir la cantidad de componentes a conservar según la varianza explicada, se suele utilizar la regla de componentes con valores propios mayores a 1 (Kaiser, 1970; Johnson & Wichern, 2002).

La ventaja de ACP es que ayudaría a generar un número reducido de variables auxiliares que pueden representar la varianza de los datos en conjunto. Otra fortaleza de este enfoque es la capacidad de incorporar información no lineal sin problemas de estimación originados por la multicolinealidad en las variables auxiliares. Como se mencionó anteriormente, los componentes principales son en principio ortogonales por tanto no existirían problemas de multicolinealidad. Así, ACP resulta en un conjunto de variables manejable y eficiente que maximizan la información contenida en el conjunto original de posibles variables auxiliares y sus elementos no lineales (Howard, Rhemtulla & Little, 2015).

De esta forma esta estrategia permite usar un conjunto de componentes principales como variables auxiliares utilizando además la información lineal y no lineal (polinomios, tales como relaciones cuadráticas) de las variables incluidas en la extracción de los componentes principales.

Para llevar a cabo este enfoque a la práctica Chesnut, Squire, Little, y Wang (2014) han desarrollado una función dentro del paquete semTools (0.4-6) (Pornprasertmanit, Miller, Schoemann & Rosseel, 2015) para el programa de código abierto R (R Core Team, 2014) llamada Quark. El objetivo de esta función es hacer este método fácil y comprensible para usuarios de R (R Core Team, 2014) basado en los resultados obtenidos por Howard, Rhemtulla y Little (2015) en sus estudio de simulación.

La función Quark brinda a los y las investigadoras los componentes principales que pueden ser utilizados para imputar los valores faltantes. La información extraída son los puntajes de componentes principales que representan los datos. Sin embargo, antes de extraer los componentes principales, la base de datos original es imputada en el caso de que existan valores perdidos (se genera un conjunto de datos con valores perdidos sustituidos). Posteriormente, los valores de componentes principales para cada sujeto son salvados y combinados con la base de datos original para ser utilizados como variables auxiliares en el proceso de imputación de datos.

Como puede apreciarse la función Quark representa una opción práctica para llevar acabo el enfoque de componentes principales como variables auxiliares. No obstante, el método aún está en constante prueba y son necesarios más estudios de simulación para determinar el número ideal de componentes principales a ser incluidos en IM y FMIL para facilitar la eficiencia y disminuir el sesgo de estimación. Asimismo, es necesario someter a prueba este proceso en diferentes escenarios aun no simulados en la literatura disponible tales como modelos multinivel (Howard, Rhemtulla & Little, 2015).

Diseño de datos perdidos para la implementación de instrumentos en psicología

Hasta el momento se ha abordado el tema de la presencia de valores perdidos como un hecho común y a veces no controlable en la labor de medir un constructo o conducta. Asimismo, la perspectiva más frecuente es creer que los datos perdidos deben ser evitados en favor de tener toda la información completa y disponible para realizar nuestras estimaciones. No obstante, no siempre los valores perdidos son una situación indeseable o negativa. Es posible diseñar una investigación o implementación de varios instrumentos de medición contemplando la presencia de valores perdidos como una estrategia para ahorrar tiempo y dinero (Little et al. 2014; Enders, 2010).

Los diseños de datos perdidos han sido ampliamente recomendados durante décadas como forma eficiente para reducir los costos, mejorar la calidad de los datos y mantener el poder estadístico para detectar los posibles efectos (Popham, 1993; Shoemaker, 1973; Sirotnik, 1974). Investigadores en el tratamiento de datos perdidos han argumentado que este tipo de diseños pueden también mejorar la validez de un instrumento en muchas circunstancias donde factores como la fatiga, dificultad o reactividad frente al test pueden ser una amenaza a la validez de una medida (Enders 2010; Graham, 2009, 2012; Little et al., 2014; van Buuren, 2012).

Al planificar la presencia de datos perdidos se puede asegurar el cumplimiento del supuesto de datos perdidos completamente al azar o MCAR. Como se mencionó anteriormente, cuando los datos perdidos obedecen al mecanismo de MCAR, no se introduce sesgo en la estimación de los parámetros del modelo, no obstante el poder estadístico puede ciertamente, verse impactado. De esta forma, al sustituir los datos perdidos, las estimaciones no son diferentes de los valores que hubiera sido estimados si no existiera valores perdidos (Graham, 2009; Little et al., 2014).

Con la ayuda de los métodos modernos para la recuperación de datos (IM y FMIL) es posible recobrar el poder estadístico que puede perderse y reducir

el sesgo las estimaciones. Para la utilización de IM y de FIML, el cumplimiento de MCAR es ideal para la recuperación de valores debido a que su impacto sobre la naturaleza de los datos es menor.

Existen diferentes diseños de datos perdidos (ver Graham, 2006) que pueden ser utilizados en la medición en psicología y educación, no obstante el objetivo del presente artículo es describir las principales características de los protocolos multi formulario y el diseño de datos perdidos para dos medidas.

Protocolos multi formulario

Tal como menciona Little et al. (2014) el diseño multi formulario más simple sería el diseño de tres formularios. En este protocolo se crean tres diferentes formularios y éstos se asignan de manera aleatoria a los participantes. El objetivo del diseño de tres formularios es asignar ítems en cuatro diferentes bloques o sets, los cuales son designados con las letras X, A, B y C. El bloque X contiene ítems que serán administrados a todos y todas las participantes. Los restantes bloques de ítems A, B y C son pareados para crear los distintos formularios: X+A+B, X+A+C y X+B+C. Esto significa, que uno de los bloques de ítems es intencionalmente omitido. En este protocolo, cada formulario contiene el 75% de los ítems del protocolo completo, asimismo más bloques de ítems pueden ser añadidos para crear formularios que tenga menos cantidad de ítems. Así, los diseños multi-formulario puede ser diseñados para generar formularios con alrededor del 40% de los ítems de la batería completa de ítems, esto quiere decir, que cada formulario podría tener hasta un 60% de datos perdidos (Raghunathan & Grizzle, 1995).

En el diseño de tres formularios se garantiza el cumplimiento del supuesto MCAR ya que cada uno de los tres formularios es asignado aleatoriamente a cada participante. La Tabla 1 muestra un esquema de los patrones de datos completos y datos perdidos esperados al implementar este diseño, de esta forma; se perdería de un 25% a un 30% de datos dependiendo del número de ítems asignados a los bloques A, B y C (Little et al., 2014).

Para utilizar este diseño es también importante tomar en cuenta varias recomendaciones, en primera instancia es valioso incluir las variables sociodemográficas y las variables que puedan predecir el patrón de datos perdidos MAR en el bloque X, el cual será administrado a todos los participantes. Asimismo, dado que puede haber otros valores perdidos no contemplados en el diseño es recomendable incluir al menos un ítem de cada constructo en el bloque X. Este ítem debería ser el mejor indicador del constructo, por ejemplo el ítem con la mejor carga factorial en un análisis de factores confirmatorio. Los restantes reactivos que representan el constructo pueden ser distribuidos de manera equitativa en los restantes bloques A, B y C (Little et al., 2014).

La clave de este diseño subyace en la correlación entre los bloques, conforme más relacionados están los bloques, más eficiente será la recuperación de la información lo cual conlleva mayor poder estadístico y proporciones de cobertura elevadas

Tabla 1

Esquema de un diseño de datos perdidos con tres formularios

Formulario	Bloque común X	Bloque A	Bloque B	Bloque C
1	25% de los items	25% de los items	25% de los items	Perdidos
2	25% de los items	25% de los items	Perdidos	25% de los items
3	25% de los items	Perdidos	25% of items	25% de los items

Nota. Las proporciones de los ítems deben ser las señaladas.

Adaptado a partir de Little et al. (2014).

cuando los datos sean analizados (Little et al., 2014; Graham, 2006).

Los diseños multi-formulario son especialmente adecuados para estudios con muestras numerosas que serán analizadas en el contexto del modelado de ecuaciones estructurales. Según estudios de simulación, este tipo de diseños requiere muestras de al menos 180 o más participantes para sostener la cobertura y convergencia adecuadas (Jia, Moore, Kinai, Crowe, Schoemann, & Little, 2014), cantidad adecuada para la estimación de modelos de ecuaciones estructurales (Little, 2013). Asimismo, ha demostrado ser una opción apropiada para diseños longitudinales como el análisis de crecimiento latente (Rhemtulla, Jia, Wu & Little, 2014)

Diseño de datos perdidos para dos medidas

En la planificación de datos perdidos es posible además de aleatorizar la presentación de los ítems, asignar al azar la implementación de instrumentos o medidas. El diseño de datos perdidos para dos medidas contempla el uso de dos medidas: una medida económica en tiempo y dinero, con baja validez y una segunda medida, costosa en tiempo y dinero con la las más elevada validez y confiabilidad posible para representar el mismo constructo (Rhemtulla & Little, 2012). La idea principal en este protocolo es asignar aleatoriamente la utilización de la medida costosa entre los y las participantes, mientras la medida menos costosa y más "ruidosa" psicométricamente sería completada por la totalidad de los y las participantes.

Para la ejecución de este diseño es importante tener en cuenta ciertos criterios tales como: (a) la medida menos costosa es una medida sistemáticamente sesgada así que es probable que este instrumento pueda representar también otros constructos, por el contrario (b) la medida más costosa debe ser una medida no sesgada, o si lo está, debe serlo en menor grado que la medida menos costosa para poder realizar la recuperación de datos de la manera más apropiada, además lo más importante es que (c) ambas medidas midan el mismo constructo, sin embargo ambas escalas pueden representar el constructo con diferentes unidades de medida o en diferentes escalas, y por ultimo; d) la investigación debe centrar sus hipótesis a nivel grupal (Rhemtulla & Little, 2012; Little et al., 2014).

Este enfoque es idóneo en el contexto del modelado de ecuaciones estructurales ya que el análisis de variables latentes permitiría corregir el sesgo de la medida menos costosa en vez de realizar alguna corrección según el sesgo previo análisis del modelo. Debido a esto, el número de participantes debe ser suficientemente numeroso aunque no se han realizado estudios específicos acerca del tamaño de la muestra (Rhemtulla & Little, 2012). Por lo general, una muestra mínima de 125 casos completos serán suficientes para realizar estimaciones de covarianza estables para trazar el modelo de ecuaciones estructurales (Rhemtulla & Little, 2012).

La principal ventaja de este diseño es la posibilidad de obtener un modelo con mayor poder estadístico, que si implementáramos solamente la medida costosa, así mismo garantiza mayor validez de constructo en comparación a un estudio que solo incluyera la medida menos costosa. Esto es así, debido a que la medida más costosa y confiable puede ser usada para modelar el sesgo asociado con la medida menos costosa, afirmación que encuentra respaldo en estudios de simulación previos donde se ha hallado que este diseño produce errores estándar reducidos y cantidades de muestra altamente efectivos para someter a prueba los parámetros del estudio (Graham, 2006).

El diseño de datos perdidos para dos medidas es en la actualidad uno de los métodos más utilizados y poderosos, con bajo costo económico y es ideal para la recolección de datos de estudios con muestras numerosas controlando el efecto de la fatiga de los y las participantes (Little et al., 2014). Además, ha demostrado ser una estrategia eficiente en diseños longitudinales disminuyendo el sesgo en las estimaciones de los errores estándar con una eficiencia aceptable (Garnier-Villarreal, Rhemtulla & Little, 2014)

Poder estadístico y tamaño de la muestra en diseños con datos perdidos

La pérdida de poder estadístico es una de las dificultades asociadas al diseño de datos perdidos. Según Enders (2010), la pérdida de poder en estos diseños no es necesariamente proporcional a la disminución en el tamaño de la muestra, por el contrario es más dependiente de la correlación entre las medidas o entre los formularios. Esta característica hace difícil obtener estimaciones precisas del poder estadístico con las estrategias de análisis comunes. No obstante, existen diferentes enfoques para estimar el poder estadístico a priori suponiendo la presencia de datos perdidos.

La estimación del poder estadístico es aún más complejo si se realiza en el contexto del modelado de ecuaciones estructurales donde un solo modelo puede tener una cantidad numerosa de parámetros, medias, intercepto, varianzas y covarianzas. Debido a esto, es también difícil estimar el tamaño de la muestra necesaria tomando en cuenta la presencia de valores perdidos. Asimismo, en un modelo de ecuaciones estructurales, el poder estadístico de un parámetro puede estar relacionado con la estimación de otro parámetro (Davey & Savla, 2009).

Existen esfuerzos previos para realizar estimaciones del tamaño muestral con datos perdidos (Muthén & Muthén, 2002; Mooijaart, 2003; Yuan & Hayashi, 2003) y cada investigación ha propuesto diferentes alcances al problema. Sin embargo ninguno de estos estudios se centra en la estimación específica del tamaño de la muestra en el diseño de investigaciones con datos perdidos.

Una de las estrategias más utilizadas para la estimación del poder y tamaño de la muestra en ecuaciones estructurales son las simulaciones Monte Carlo (Muthén & Muthén, 2002; Enders, 2010). El propósito de este enfoque es extraer un número grande de muestras aleatorias (e.g., 1000) de una población definida por la hipótesis alternativa y estimar el modelo planeado en las distintas muestra. De esta forma, el poder estadístico puede ser calculado como la proporción de muestras que rechazan la hipótesis nula (Schoemann et al., 2014).

A pesar de las dificultades señaladas para el análisis de poder estadístico en los diseños de datos perdidos, existe un estudio de simulación que busca responder a la pregunta del número mínimo necesario de participantes para la ejecución de un diseño de datos perdidos con tres formularios. Esta investigación realizada por Jia et al. (2014) demuestra que un mínimo de 90 participantes puede ser suficiente para estimar un modelo de factores confirmatorio en un diseño transversal, cuando se utiliza FMIL y bajo las condiciones especificadas por los autores. Asimismo, si se desea trazar un modelo de factores confirmatorio con dos puntos de medición la muestra debería estar integrado por un mínimo de130 participantes cuando se utiliza FMIL y 175 personas cuando se implementa IM.

Si bien estudios como los de Jia et al (2014) son rigurosos, es poco probable realizar simulaciones en todas las condiciones y combinaciones posibles en el modelado de ecuaciones estructurales para estimar el poder estadístico o el tamaño de la muestra necesario para ejecutar diseños de valores perdidos. Para sortear esta desventaja, existen varios programas estadísticos que pueden ser utilizados para realizar simulaciones de Monte Carlo (Enders, 2010).

Uno de los más populares es MPLUS (Muthén & Muthén, 1998-2013), sin embargo existe otra alternativa gratuita y de acceso abierto llamado simsem (Pornprasertmanit, Miller, & Schoemann, 2013), éste es un paquete que puede ser descargado en el entorno del programa R (R Core Team, 2014). La finalidad de este paquete es facilitar el uso de grandes simulaciones de Monte Carlo en el contexto del modelado de ecuaciones estructurales. Actualmente, simsem es capaz de generar datos y utilizar el paquete lavaan (Rosseel, 2012) o el paquete OpenMx (Boker, et al., 2012) para analizar los datos simulados. Además, simsem fue especialmente diseñado para simular valores perdidos cumpliendo los supuestos de MCAR, MAR y diseños de datos perdidos específicos (Schoemann et al., 2014).

Conclusión

Es evidente como la presencia de datos perdidos no significa una amenaza para evaluación de una medida o la ejecución de una investigación. Siempre que se cumpla el supuesto de datos perdidos MCAR o el más común MAR. Asimismo, se expuso la fortaleza de las variables auxiliares como una solución para la recuperación de la información bajo del supuesto de MAR.

Las variables auxiliares pueden ser tratadas siguiendo el enfoque de componentes principales propuesto por Howard, Rhemtulla y Little (2015) no obstante, es una estrategia de reciente data y requiere de mayor investigación acerca de las posibles ventajas de los componentes principales en otros escenarios posibles y precisar la cantidad de componentes principales mínimos necesarios para mejorar las estimaciones de IM y FMIL bajo el supuesto de MAR (Howard, Rhemtulla & Little, 2015).

Aunado a lo anterior, la función Quark (Chesnut et al., 2014) supone una herramienta útil para someter a prueba el enfoque de componentes principales como variables auxiliares. Al ser una función en etapas iniciales más investigación y pruebas son requeridas para garantizar la estabilidad de la estimación (Chesnut et al., 2014) para lo cual, usuarios alrededor del mundo puede colaborar al usar esta función con sus propias bases de datos.

En esta misma línea, el paquete simsem (Pornprasertmanit et al., 2014) ofrece la ventaja de estimar el poder estadístico de los diseños de datos perdidos para complementar la fase de desarrollo de un diseño de este tipo. Su principal ventaja es su capacidad para simular modelos complejos con datos perdidos y su interacción con otros paquetes para el análisis de la información simulada (Schoemann et al., 2014).

Es necesario acuñar mayor evidencia acerca de diseños con datos perdidos con mayor cantidad de datos perdidos como lo puede ser un diseño de datos perdidos con 10 formularios, especialmente en diseños longitudinales. Sin embargo no existe evidencia para dudar de su efectividad y su utilidad para la implementación de instrumentos de auto reporte (Little et al. 2014).

En resumen, es importante cambiar la actitud de los y las investigadoras frente a enfoques modernos para el manejo de datos perdidos (Little, 2013). Estas herramientas representan una oportunidad de mejorar la calidad de los análisis estadísticos realizados y los modelos trazados en psicología y educación. Teniendo en cuenta que las herramientas computacionales lo permiten, es posible cambiar el acercamiento que los y las investigadores poseen hacia el tema de los valores perdidos.

Referencias

Allison, P. D. (2001). *Missing data*. Thousand Oaks: CA: Sage.

Arbuckle, J.L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5–37.

Boker, S., Neale, M., Maes, H. H., Wilde, M., Spiegel, M., Brick, T., . . Fox, J. (2011). OpenMx:Anopensour ceextendedstructuralequation modeling framework. *Psychometrika, 76*, 306–317

Chesnut, S. R., Squire, D., Little, T. D., & Wang, E. W. (2014). *Quark: An R library for preparing large datasets for multiple imputation with auxiliary variables*. [SOFTWARE ADD-ON], USA, Texas Tech University, Institute of Measurement, Methodology, and Policy (IMMAP).

Collins, L. M., Schafer, J. L., & Cam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods, 6*(4). 330-51.

Davey, A., & Savla, J. (2009). Estimating Statistical Power With Incomplete Data. *Organizational Research Methods, 12*(2), 320–346.

Enders, C. K. & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling, 8*(3), 430–457.

Enders, C. (2010). *Applied Missing Data Analysis-Methodology in Social Sciences*. New York: Guilford Press.

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. D. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development, 38*(5), 411–422.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*(1), 1–25. doi: 10.1080/00273171.2012.640589

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*, 197–218.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling , 10*, 80-100.

Graham, J. W., Taylor, B. J., & Olchowski A. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*(4), 323-343.

Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60,* 549-76.

Graham, J. W. (2012) *Missing data: Analysis and design*. New York: Springer.

Harel, O., Stratton, J., & Aseltine, R. (2011). *Designed missingness to better estimate efficacy of behavioral studies* (Technical Report 11-15). Storrs, CT: Department of Statistics, University of Connecticut.

Howard, W. J., Rhemtulla, M. & Little, T. D. (2015). Using Principal Components as Auxiliary Variables in Missing Data Estimation. *Multivariate Behavioral Research, 50*(3), 285-299.

Jia, F., Moore, E. W. G., Kinai, R., Crowe, K. S., Schoemann, A. M., & Little, T. D. (2014). Planned missing data design on small sample size: How small is too small? *International Journal of Behavioral Development, 38*(5), 435-452.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-415.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling*. (3rd edition). New York: The Guilford Press.

Little, R. K. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

Little, T.D., Jorgensen, T.D., Lang, K.M., & Moore, E.W. (2014).On the Joys of Missing Data. *Journal of Pediatric Psychology, 39*(2), 151-162.

Mooijaart, A. (2003). Estimating the statistical power in small samples by empirical distributions. En: H. Yanai, A. Okada, K. Shigemasu,Y. Kano & J.J. Meulman (Eds.), *New developments in psychometrics* (pp. 149-156). Japan: Springer-Verlag.

Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen.

Muthén, L. K., & Muthén, B.O. (2002). How to use a Monte Carlo Study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2014). *SIMSEM: Simulated structural equation modeling. R package version 0.5–8*. Recuperado de: http://www.simsem.org

Pornprasertmanit, S., Miller, P., Schoemann, A. & Rosseel, Y. (2015). *semTools: Useful tools for structural equation modeling. R package version 0.4–6*.

Recuperado de: http://CRAN.R-project.org/package=semTools.

Popham, W. J. (1993). Circumventing the high costs of authentic assessment. *Phi Delta Kappan, 74*(6), 470-473.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de http://www.R-project.org/.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association, 90*, 54–63.

Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (2014). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development, 38*(5), 423–434.

Rhemtulla, M. & Little, T. (2012). Tools of the Trade: Planned Missing Data Designs for Research in Cognitive Development. *Journal of Cognitive Development, 13*(4), 425-438

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. Disponible en: http://lavaan.ugent.be/

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). A*nalysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147–177.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger.

Schoemann, A. M., Miller, P., Pornprasertmanit, S., & Wu, W. (2014). Using Monte Carlo simulations to determine power and sample size for planned missing designs. *International Journal of Behavioral Development, 38*(5), 471-479.

Sirontnik, K.A. (1974). Introduction to matrix sampling for the practitioner. In W.J. Pophan (ed.), *Evaluation in Education* (pp.453-529). Berkeley, CA: McCurtchau Publishing Corp.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

Yuan, K., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 56*, 93–110.

Recibido: 20 de Mayo de 2015
Aceptado: 14 de Setiembre de 2015