Psicothema 2012. Vol. 24, nº 1, pp. 167-175
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Imputación de datos perdidos en las evaluaciones diagnósticas educativas

Rubén Fernández-Alonso[1], Javier Suárez-Álvarez[2] y José Muñiz[2]
[1] Consejería de Educación y Universidades del Gobierno del Principado de Asturias y [2] Universidad de Oviedo

En la evaluación diagnóstica de sistemas educativos se utilizan habitualmente autoinformes para recoger datos de carácter tanto cognitivo como oréctico. Es muy frecuente que por distintas razones en estos autoinformes falten algunos de los datos del alumnado. El objetivo del presente trabajo es comparar el funcionamiento de diferentes métodos de imputación de datos perdidos en el contexto de la evaluación de sistemas educativos. Sobre una base de datos de 5.000 sujetos se simularon 72 condiciones: tres tamaños de pérdida de datos, tres mecanismos de pérdida y ocho métodos de imputación de los datos perdidos. La cuantía de las pérdidas se establecieron en un 5, 10 y 20%. Los mecanismos de pérdida fijados fueron: aleatoria, moderadamente condicionada y fuertemente condicionada. Los ocho métodos de imputación utilizados fueron: eliminación, reemplazo por la media de la escala, por la media del ítem, por la media del sujeto, por la media del sujeto corregida, regresión múltiple e imputación por el algoritmo Esperanza-Maximización (EM) con y sin variables auxiliares. Los resultados indican que la recuperación de los datos es más precisa cuando se emplea una combinación adecuada de diferentes métodos de recuperación de los datos perdidos. Cuando se trata de un caso incompleto funciona muy bien la media del sujeto, mientras que para datos completamente perdidos es recomendable la imputación múltiple con el algoritmo EM. El uso de esta combinación resulta especialmente recomendable cuando la pérdida de datos es mayor y su mecanismo de pérdida está más condicionado. Finalmente, se discuten los resultados y se comentan algunas líneas futuras de investigación que se abren a partir de los resultados obtenidos.

Imputation methods for missing data in educational diagnostic evaluation. In the diagnostic evaluation of educational systems, self-reports are commonly used to collect data, both cognitive and orectic. For various reasons, in these self-reports, some of the students' data are frequently missing. The main goal of this research is to compare the performance of different imputation methods for missing data in the context of the evaluation of educational systems. On an empirical database of 5,000 subjects, 72 conditions were simulated: three levels of missing data, three types of loss mechanisms, and eight methods of imputation. The levels of missing data were 5%, 10%, and 20%. The loss mechanisms were set at: Missing completely at random, moderately conditioned, and strongly conditioned. The eight imputation methods used were: listwise deletion, replacement by the mean of the scale, by the item mean, the subject mean, the corrected subject mean, multiple regression, and Expectation-Maximization (EM) algorithm, with and without auxiliary variables. The results indicate that the recovery of the data is more accurate when using an appropriate combination of different methods of recovering lost data. When a case is incomplete, the mean of the subject works very well, whereas for completely lost data, multiple imputation with the EM algorithm is recommended. The use of this combination is especially recommended when data loss is greater and its loss mechanism is more conditioned. Lastly, the results are discussed, and some future lines of research are analyzed.

Los programas de evaluación de los sistemas educativos van dirigidos a poblaciones muy amplias de estudiantes, y en la recogida de datos emplean como metodología primaria, aunque no exclusiva, el autoinforme, ya sea mediante tests para medir variables cognitivas, o a través de cuestionarios para recoger información del contexto educativo y de variables orécticas (Fernández-Alonso y Muñiz, 2011; Ministerio de Educación, 2009; Mullis, Martin, Kennedy, Trong y Sainsbury, 2009; Mullis, Martin, Ruddock, O'Sullivan y Preuschoff, 2009; National Assessment Governing Board, 2003; OECD, 2009). En este contexto el autoinforme es un procedimiento eficiente y económico, aunque presenta ciertas limitaciones, tanto de carácter epistemológico como psicométricas, ejemplo de las cuales es la aparición de datos perdidos (Bokossa y Huang, 2001; Willms y Smith, 2006).

El tratamiento estadístico de datos perdidos, aún siendo un tema clásico, resulta de plena actualidad en la medida en que los

Fecha recepción: 22-12-10 • Fecha aceptación: 12-5-11
Correspondencia: José Muñiz
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: jmuniz@uniovi.es

avances informáticos permiten operativizar nuevas formas de dar solución a este tipo de omisiones. En algunos casos no es suficiente con eliminar directamente estos datos, pues se pierde potencia de la prueba y se renuncia a explotar toda la información disponible (Botella, 2002). Los trabajos más clásicos plantean algunas soluciones sencillas como el reemplazo del valor perdido por la media del ítem, la media de la escala o la media del sujeto (Downey y King, 1998; Graham, 2009). Sin embargo, la rapidez de procesamiento que proporciona la informática ha permitido el desarrollo de otros modelos que aún siendo menos parsimoniosos mejoran la estimación de los parámetros estadísticos. Así, por ejemplo, el uso de la regresión (Shrive, Stuart, Quan y Ghali, 2006; Van Ginkel, Sijtsma, Van der Ark y Vermunt, 2010), o algoritmos basados en la máxima verosimilitud (Enders, 2004; Graham, 2009; San Luis, Hernández y Ramírez, 1997), consiguen una importante representatividad en la literatura psicométrica. Otros autores proponen el uso de la probabilidad inversa y estimadores doblemente robustos para el análisis de datos incompletos (Vansteelandt, Carpenter y Kenward, 2010), así como imputaciones multivariantes (Van Buuren, 2010). Pero la tendencia hegemónica, desde Rubin (1987) hasta la actualidad (Graham, 2009; Shrive, Stuart, Quan y Ghali, 2006; Van Ginkel, Sijtsma, Van der Ark y Vermunt, 2010), se basa en la utilización de múltiples imputaciones que permitan ganar precisión estadística.

Nuestro trabajo trata de explorar la eficacia de los métodos de imputación en el contexto específico y muy peculiar de la evaluación de los sistemas educativos. Los datos utilizados provienen de la evaluación de diagnóstico educativo del Principado de Asturias, iniciada en el año 2005 (Gobierno del Principado de Asturias, 2011). Las razones por las que se pierden datos en una evaluación educativa son muy variadas (Yamamoto, 2001), lo que obliga a un análisis particular de las pérdidas de cada estudio. A lo largo de estos años la evaluación asturiana ha ido acumulando información sobre el tamaño y características de sus pérdidas de datos. Estos análisis indican que, por término medio, la proporción de datos perdidos no es muy grande, encontrándose entre el 5 y el 10%. Los análisis también ponen de manifiesto que la pérdida de datos no se ajusta a un mecanismo completamente aleatorio, según la definición que del mismo hacen Little y Rubin (1987), sino que el grupo de estudiantes con valores perdidos presenta sistemáticamente un resultado más bajo en las pruebas de desempeño académico y en otros índices vinculados a dicho desempeño. Todas estas evidencias no han sido aún publicadas, por ello, a continuación se hará la justificación oportuna. Para ilustrar estas dos afirmaciones se muestran los datos de la Evaluación de Diagnóstico Asturias 2010, que serán la base sobre la que se desarrollará el presente trabajo. En concreto nos centraremos en el análisis de los valores perdidos en dos de las variables centrales de la Evaluación de Diagnóstico: el autoconcepto académico y el rendimiento en matemáticas. Entre las preguntas que respondieron los estudiantes de la evaluación Asturias 2010 había cinco destinadas a medir el autoconcepto académico. La tabla 1 muestra la distribución de casos según el patrón de respuestas perdidas.

Tabla 1
Tipos de datos en la Escala de Autoconcepto Académico

	N
Datos completos	6785
Datos incompletos o parcialmente perdidos	487
Datos totalmente perdidos	330
Total	7602

Se observa que el 89,25% del total son datos completos, es decir, los estudiantes han respondido a los cinco ítems. El 4,34% de los casos están totalmente perdidos y el 6,41% tienen datos incompletos, es decir, solo se dispone de información en alguno de los ítems.

Por otra parte, la tabla 2 muestra que el grupo de estudiantes con datos completos presenta el mejor promedio tanto en matemáticas como en autoconcepto. En todas las comparaciones las diferencias son estadísticamente significativas a favor de este grupo.

En este contexto de evaluación de sistemas educativos, el objetivo de la presente investigación es el de indagar en el comportamiento de ocho métodos de imputación en función de la cuantía de la pérdida de los datos y del mecanismo de pérdida. Para alcanzar este objetivo trabajaremos generando datos simulados a partir de la matriz de datos reales de la Evaluación de Diagnóstico de Asturias 2010.

Método

Participantes

El presente estudio supone una explotación especial de la base de datos de la Evaluación de Diagnóstico Asturias 2010. Dicha base está compuesta por 7.602 estudiantes de cuarto curso de Educación Primaria que respondieron a una prueba de competencia Matemática y de Conocimiento e interacción con el mundo físico, y a un cuestionario de contexto educativo. De la base de datos original se seleccionó una muestra aleatoria de 5.000 casos con datos completos en las variables de interés: una Escala de *Autoconcepto Académico* de cinco ítems y una prueba de *Competencia en Matemáticas* de 96 ítems. Esta muestra final de 5.000 casos tiene una distribución de edad con media de 9,86 y desviación típica de 0,511, a su vez, cuenta con un 50,4% de niños y un 49,6% de niñas.

Instrumentos

Escala de autoconcepto académico

El autoconcepto se midió con una escala de cinco ítems tipo Likert con cuatro categorías de respuesta: nunca o casi nunca, a

Tabla 2
Media y error típico en Matemáticas y Autoconcepto según el tipo de datos en la Escala de Autoconcepto

Tipo de datos en autoconcepto	Matemáticas		Autoconcepto	
	N	Media (E.T.)	N	Media (E.T.)
Datos completos	6554	502.1 (1.23)	6785	2.08 (0.0086)
Datos incompletos	455	473.6 (4.57)	487	1.84 (0.0328)
Datos perdidos	56	464.7 (13.4)	330	n.a.
Total	7065	500.0 (1.19)	7602	2.06 (0.0084)

n.a.: no aplica: para el alumnado con datos totalmente perdidos en autoconcepto no es posible calcular ningún estadístico en dicha variable
E.T.: Error típico de la media

veces, a menudo y siempre o casi siempre. En la codificación de las respuestas se empleó una escala de 0 a 3 puntos.

La elección de la escala de autoconcepto se basa en dos razones. En primer lugar por sus propiedades psicométricas, ya que presenta una alta consistencia interna (α= 0,85) a pesar de ser una prueba muy corta (cinco ítems), y es esencialmente unidimensional (el primer factor explica el 62,7% de la varianza total de los ítems). La segunda razón es que, en las evaluaciones del sistema educativo asturiano, el autoconcepto se ha mostrado como la variable individual mejor conectada con los resultados educativos (Gobierno del Principado de Asturias, 2008, 2010, 2011).

Prueba de Competencia Matemática

Como variable criterio se ha utilizado la Competencia Matemática, la cual ha sido estimada mediante diferentes modelos matemáticos derivados de la Teoría de Respuesta al Ítem. Se trata de una prueba compuesta por un total de 96 ítems repartidos entre ítems de elección múltiple, respuesta abierta corta y respuesta abierta construida. En lo que respecta a sus propiedades psicométricas se trata de una prueba fiable (α= 0,86) y válida en cuanto a su contenido (un grupo de expertos en el tema elaboraron la prueba a partir de las especificaciones de la Consejería). Dicha competencia se expresa en una escala cuya media es de 500 puntos y la desviación típica 100. Para más detalles véase, por ejemplo, Gobierno del Principado de Asturias (2008, 2010, 2011).

Procedimiento

Las direcciones de los centros educativos han sido las responsables de la gestión de todo el proceso: información a la comunidad educativa de la finalidad e importancia de la prueba, asignación de las personas responsables de la aplicación, custodia de las pruebas hasta el momento de su aplicación, así como de realizar el control interno.

Las pruebas de la Evaluación de Diagnóstico Asturias 2010 fueron aplicadas en todos los casos por el profesorado que la Dirección de cada centro docente determinó. Las pautas específicas para la aplicación de las pruebas se proporcionaron en el Protocolo para el desarrollo de la evaluación de diagnóstico en la parte correspondiente a los cuadernillos de cada competencia (Gobierno del Principado de Asturias, 2011).

Diseño

En las simulaciones realizadas se manipulan tres parámetros: cuantía de la pérdida de datos (tres niveles), mecanismo de pérdida (tres niveles) y métodos de imputación (ocho métodos). Se trata por tanto de un diseño cruzado de $3\times3\times8$, en total se simularán 72 condiciones. Se tomará como línea base la matriz de datos completos.

Cuantía de las pérdidas

En función del tamaño de la pérdida se han establecido tres condiciones:

a) Pérdida del 5% de las respuestas. Dado que la base de datos tiene 5.000 sujetos en esta condición se han eliminado de cada uno de los ítems 250 respuestas. Esta eliminación es independiente en cada ítem, de tal forma que en algunos sujetos se pierde el dato en uno de los cinco ítems y en otros casos llegan a perderse todas las respuestas. Esto hace que, en la práctica, los casos perdidos en el procedimiento *listwise deletion* sean bastantes más del 5%.
b) Pérdida del 10%. En cada ítem se eliminan las respuestas de 500 sujetos. El modo en que se eliminan los casos es el comentado en el párrafo anterior y, por tanto, la advertencia sobre el porcentaje total de casos incompletos (los eliminados en el *listwise deletion*) también es aplicable aquí.
c) Pérdidas del 20%. En cada ítem se eliminan 1.000 respuestas con el mismo procedimiento y reservas apuntadas en las dos condiciones anteriores.

La elección de estos porcentajes se justifica en el análisis del tamaño de la pérdida de datos presentada en la introducción. Las condiciones de pérdida del 5 y del 10% suponen los límites razonables del porcentaje de pérdida de estos estudios, mientras que la pérdida del 20% se incluye como una condición extrema y menos probable en la evaluación diagnóstica aplicada.

Mecanismo de pérdida

El segundo criterio para simular la pérdida de los datos es el mecanismo de pérdida, considerando tres condiciones:

a) Pérdida completamente aleatoria (MCAR: *Missing completely at random*). En este caso las respuestas a eliminar en cada ítem (250, 500 o 1.000 según el tamaño de pérdida) fueron determinadas mediante la selección aleatoria del número de casos equivalentes al tamaño de la pérdida deseada.
b) Pérdida de datos condicionada moderadamente (NMAR 60/40: *Not missing at random*), a la que también nos referiremos como pérdida intencional 60/40. Para explicar este procedimiento se ejemplifica una pérdida concreta. Supongamos que se quieren eliminar 250 casos en el ítem 2 de la escala: *«saco buenas notas»*. Para ello se crean dos grupos, el grupo A compuesto por los casos que opinan que nunca o solo a veces sacan buenas notas, y el grupo B conformado por los sujetos que afirman que a menudo o siempre sacan buenas notas. En la selección de casos perdidos 6 de cada 10 estudiantes son del grupo A y los 4 restantes del grupo B. Así que cuando es necesario eliminar 250 casos, 150 se perderán en el grupo A y 100 en el grupo B, lo cual se hará de forma aleatoria dentro de cada grupo.
c) Pérdida de datos condicionada fuertemente (NMAR 80/20), que también será denominada como «pérdida intencional 80/20». En este caso 8 de cada 10 casos perdidos se seleccionan del grupo A y los otros 2 del grupo B, de forma aleatoria dentro de cada grupo.

La simulación de los mecanismos de pérdida intencional se basa en el análisis de los datos perdidos mostrado en la introducción, donde se muestra que los datos no se pierden aleatoriamente, sino que los estudiantes con datos perdidos presentan un autoconcepto más bajo. La tabla 3 resume las 9 condiciones de pérdida de datos simuladas (3 tamaños $\times$ 3 mecanismos de pérdida). Dentro de cada condición se muestra también el resultado del Test de Little (1988), cuyos datos confirman que los datos de las dos columnas de la derecha han sido eliminados condicionalmente.

Tabla 3
Modo de selección de casos perdidos en función del tamaño y mecanismo de pérdida y resultado del test de Little para cada condición de pérdida

Tamaño de la pérdida	Mecanismo de pérdida		
	Aleatoria (MCAR)	**Condiciona Moderada (NMAR 60/40)**	**Condiciona Fuerte (NMAR 80/20)**
5%	250 casos al azar	150 casos al azar de la mitad negativa 100 casos al azar de la mitad positiva	200 casos al azar de la mitad negativa 50 casos al azar de la mitad positiva
(250 casos perdidos por ítem)	Test MCAR: χ^2= 44.0; gl.=54 (p= 0.833)	Test MCAR: χ^2= 547; gl.= 86 (p<0.000)	Test MCAR: χ^2= 662; gl.= 69 (p<0.000)
10%	500 casos al azar	300 casos al azar de la mitad negativa 200 casos al azar de la mitad positiva	400 casos al azar de la mitad negativa 100 casos al azar de la mitad positiva
(500 casos perdidos por ítem)	Test MCAR: χ^2= 69.5; gl.=71 (p= 0.528)	Test MCAR: χ^2= 822; gl.=99 (p<0.000)	Test MCAR: χ^2= 1219; gl.= 75 (p<0.000)
20%	1000 casos al azar	600 casos al azar de la mitad negativa 400 casos al azar de la mitad positiva	800 casos al azar de la mitad negativa 200 casos al azar de la mitad positiva
(1.000 casos perdidos por ítem)	Test MCAR: χ^2= 73.7; gl.=75 (p= 0.520)	Test MCAR: χ^2= 1374; gl.= 105 (p<0.000)	Test MCAR: χ^2= 1441; gl.= 75 (p<0.000)

Tabla 4
Diferencia de medias en Matemáticas y Autoconcepto según el patrón y el mecanismo de pérdida

Mecanismo de pérdida	Patrón de pérdida	Matemáticas		Autoconcepto académico	
		Media (s.e.)	Diferencia	Media (s.e.)	Diferencia
MCAR	Datos completos	506 (1.80)	●	2.09 (0.01)	●
	Datos perdidos	507 (2.26)		2.13 (0.02)	
NMAR 60/40	Datos completos	517 (1.76)	▲	2.28 (0.01)	▲
	Datos perdidos	490 (2.29)		1.85 (0.02)	
NMAR 80/20	Datos completos	522 (1.72)	▲	2.37 (0.01)	▲
	Datos perdidos	479 (2.30)		1.66 (0.02)	

●: Sin diferencias estadísticamente significativas (p<0,01)
▲ : Con diferencias estadísticamente significativas (p<0,01)

Para confirmar las afirmaciones del párrafo anterior se muestran los resultados de la tabla 4. En ella se resume la media en matemáticas y autoconcepto en función del patrón de pérdida y el mecanismo de la misma. Cuando la pérdida es MCAR las diferencias entre los dos grupos no son estadísticamente significativas. En cambio, con pérdidas NMAR las diferencias en los promedios de matemáticas y autoconcepto a favor del grupo con datos completos son estadísticamente significativas.

Métodos de imputación

La literatura sobre métodos de imputación es ciertamente abundante (Graham, 2009; Howell, 2007; Little, 1992), aquí hemos elegido ocho métodos realistas, con posibilidades de ser aplicados en la práctica de la evaluación diagnóstica de sistemas educativos. Se han empleado métodos, como el análisis de datos completos, que se consideran las aproximaciones más elementales al análisis de datos perdidos, también los considerados clásicos (Graham, 2009; Howell, 2007). Todos ellos se basan en reemplazar el valor perdido por alguna información disponible. Dentro de este grupo se han aplicado dos procedimientos basados en usar la información disponible sobre la escala completa o sobre los ítems individuales que la componen, y otros dos procedimientos que consisten en recuperar los datos faltantes basándose en la información proporcionada por el sujeto.

Los métodos de reemplazo descritos hasta ahora, especialmente la sustitución del valor perdido por información de la escala, han sido objeto de críticas. Las mismas indican que dicha sustitución no aporta información nueva y en cambio reduce el error típico del estimador (Cohen, Cohen, West y Aiken, 2003). Una posibilidad es sustituir el valor perdido empleando un modelo de regresión y adicionalmente crear una variable dummy D que es igual a 1 cuando el dato está perdido y 0 en cualquier otro caso (Cohen y Cohen, 1985). Si bien la mayoría de autores incluyen la regresión dentro de los procedimientos clásicos de recuperación de datos (Graham, 2009; Howell, 2007), su método de estimación (máxima verosimilitud) y el modo de tratar el error típico hace que en ocasiones aparezca al mismo nivel que los métodos modernos (Willms y Smith, 2006). Por estas razones, y sin querer entrar en disputas académicas, en este trabajo el modelo de regresión se sitúa a caballo entre los métodos clásicos y los modernos.

Finalmente, se emplearán dos procedimientos considerados modernos. Ambos se basan en la estimación de máxima verosimilitud, concretamente en el empleo del algoritmo de Esperanza-Maximización (EM). La lógica del mismo está bien descrita en Schafer (1999), quien señala que si los valores perdidos fuesen conocidos sería muy simple estimar los parámetros del modelo (medias, correlaciones y covarianzas), y viceversa, si se conocen estos parámetros es posible calcular los valores perdidos. El algoritmo EM es un método iterativo que calcula los valores perdidos condicionados por las estimaciones de los parámetros del modelo.

En concreto los métodos empleados fueron los siguientes:

– *Listwise deletion o pérdida de observación*. Es el procedimiento más sencillo, consiste en eliminar del análisis cualquier caso que presente al menos un dato perdido en alguno

de los cinco ítems de la escala autoconcepto. En puridad, no es un método de imputación de datos perdidos, aquí se utiliza para comparar el efecto que operar solo con casos completos tiene sobre las inferencias que se realizan en el análisis.

- *Reemplazar el valor perdido por la media de la escala*. En este caso se computa una variable de apoyo que es igual a la media de la escala, que es calculada a partir de todos los datos disponibles. Cuando un ítem aparece perdido dicho dato se reemplaza por la variable de apoyo.
- *Reemplazar el valor perdido por la media del ítem*. En nuestro análisis se crean cinco variables de apoyo, que resumen la media de cada ítem calculada a partir de las respuestas válidas. Cuando aparece un dato perdido el mismo es reemplazado por la media del ítem correspondiente.
- *Reemplazar el valor perdido por la media del sujeto*. En este caso la variable de apoyo se calcula como la media del sujeto en autoconcepto. Cuando en algún ítem aparece un dato perdido éste es reemplazado por la variable de apoyo. Este método solo se puede emplear con datos incompletos, siendo imposible su utilización cuando los datos están totalmente perdidos.
- *Reemplazar el dato perdido por la media del sujeto corregida*. Se trata de un método basado en la media del sujeto, pero que incluye un factor de corrección que tiene en cuenta la dificultad del ítem donde se ha perdido el dato y la dificultad total de la escala. Al igual que el anterior, solo se puede emplear con casos incompletos. Una buena descripción del procedimiento puede verse en Van Ginkel, Sijtsma, Van der Ark y Vermunt (2010) y Van Ginkel y Van der Ark (2005).
- *Reemplazar el dato perdido por la regresión múltiple*. A partir de las respuestas válidas a los cinco ítems se ajusta un modelo de regresión lineal múltiple por el procedimiento de máxima verosimilitud. En el mismo las respuestas a los cinco ítems del autoconcepto funcionan al tiempo como predictores y criterios. Para solucionar el problema de la infraestimación del error el módulo *Missing Value Analysis* de SPSS añade un pequeño error aleatorio en cada sustitución (Howell, 2007).
- *Algoritmo EM*. Se trata de un procedimiento de máxima verosimilitud donde cada interacción tiene dos fases (esperanza y maximización) que son las que le dan nombre: algoritmo EM. En el paso E y a partir de las respuestas válidas en los cinco ítems de autoconcepto se calculan los parámetros del modelo, esto es, las medias de los ítems y sus correlaciones y covarianzas. Con esta primera estimación de los parámetros se calculan los valores perdidos, que están por tanto condicionados tanto por el vector de respuestas como por los parámetros estimados inicialmente. En el paso M y basándose en la estimaciones del paso E se recalculan de nuevo los parámetros del modelo mediante máxima verosimilitud. En la segunda y sucesivas iteraciones se repiten las dos fases y así hasta que los datos del modelo alcanzan el criterio de convergencia.
- *Algoritmo EM con variables auxiliares*. La lógica del análisis es la misma que se acaba de describir, sin embargo en este caso el modelo no incluye solo las respuestas a los cinco ítems de autoconcepto, también se introducen tres nuevas variables altamente correlacionadas con los cinco ítems: la media del sujeto en autoconcepto, la puntuación del sujeto en un índice de motivación académica y la puntuación en matemáticas. Todas ellas se emplean como variables antecedentes vinculadas al autoconcepto académico.

Análisis de los datos

Para comparar la capacidad de los métodos de recuperación de los datos se ha calculado un abanico amplio de estadísticos, que pueden ser clasificados entre dos grupos. En primer lugar se comprueba cómo las diferentes condiciones de pérdida y tratamiento de los datos afecta a los estadísticos de posición y variabilidad de la escala. Por otro lado se verifican los estadísticos de consistencia interna y dimensionalidad. Se utilizó el programa SPSS, incluyendo su módulo *Missing Value Analysis*, el cual si bien tiene sus detractores (Von Hippel, 2004), no es menos cierto que en contextos aplicados presenta claras ventajas (Van Ginkel y Van der Ark, 2005).

Resultados

El comportamiento de los estadísticos de posición y variabilidad se muestra en la tabla 5. Como era esperable, se observa que la pérdida aleatoria recupera mejor los datos que la pérdida intencional. Esta afirmación es válida para cualquier tamaño de pérdida y método de imputación.

En pérdidas MCAR pequeñas —en torno al 5%— cualquier método de imputación parece replicar razonablemente los parámetros poblacionales. Sin embargo, en pérdidas MCAR mayores —en torno al 20%— ya se puede adelantar que los métodos basados en la información parcial de los sujetos (reemplazo por la media del sujeto con o sin corrección) y los métodos de imputación (especialmente el algoritmo EM con variables auxiliares) son los que, en conjunto, mejor recuperan los estadísticos de los datos completos.

Por su parte, bajo pérdidas MNAR todos los métodos de recuperación de datos muestran el mismo patrón: sobreestimación de la media e infraestimación de la varianza poblacional. Este doble efecto se agrava a medida que aumenta la proporción de datos perdidos y que éstos se aglutinan en el grupo de bajo autoconcepto académico. A su vez, en pérdidas NMAR el error estándar de la media tiende a infraestimarse en todos los casos salvo en el método *listwise deletion*. Sin embargo, éste es claramente el peor procedimiento bajo el mecanismo NMAR: incluso con pérdidas pequeñas (5%) la media está tan sobreestimada que se vuelve estadísticamente distinta del parámetro poblacional.

Con pérdidas NMAR el deterioro de los estadísticos de la distribución es más acusado en unos métodos de imputación que en otros. Ello hace que sea precisamente bajo las dos condiciones de pérdida intencional cuando se puede apreciar mejor la capacidad de los diferentes métodos para recuperar los parámetros originales. En pérdidas NMAR se confirma con claridad que los tres mejores métodos para recuperar los estadísticos fundamentales de la distribución son: el reemplazo por la media con y sin corrección y el algoritmo EM con variables auxiliares, presentando los dos últimos un ajuste ligeramente superior al reemplazo por la media con corrección. De hecho se aprecia que bajo la condición NMAR 60/40 y pérdida del 10% el reemplazo por la media sin corrección y el algoritmo EM replican los tres parámetros de forma casi perfecta.

En las condiciones NMAR 80/20 las diferencias de los métodos para recuperar información se vuelven más evidentes. En estas condiciones el método *listwise deletion* y los métodos basados en la información de la escala (reemplazo por la media del ítem y por la media del test) no parecen recomendables.

Tabla 5
Media, error típico de la media y varianza de la escala autoconcepto en función del mecanismo de pérdida, el tamaño y el método de recuperación de los datos

Mecanismo de pérdida	Método de imputación	Tamaño de la pérdida								
		5%			10%			20%		
		Media	S.E. Media	Var.	Media	S.E. Media	Var.	Media	S.E. Media	Var.
Datos completos (N= 5000)		**2.111**	**.0099**	**.489**	**2.111**	**.0099**	**.489**	**2.111**	**.0099**	**.489**
Aleatoria MCAR	Listwise	2.111	.0111	.482	2.094	.0129	.494	2.078	.0177	.504
	Media escala	2.111	.0094	.446	2.109	.0091	.412	2.106	.0082	.339
	Media ítem	2.111	.0094	.445	2.108	.0091	.412	2.105	.0082	.338
	Media sujeto	2.112	.0099	.491	2.112	.0100	.498	2.111	.0101	.514
	Media suj. corregida	2.112	.0099	.491	2.112	.0100	.498	2.110	.0101	.512
	Regresión	2.111	.0098	.481	2.111	.0098	.479	2.109	.0097	.468
	EM	2.111	.0098	.482	2.111	.0098	.479	2.109	.0097	.468
	EM con var. auxiliar	2.112	.0099	.491	2.112	.0100	.498	2.111	.0101	.514
Condicionada NMAR 60/40	Listwise	2.201	.0106	.435	2.279	.0113	.386	2.432	.0121	.266
	Media escala	2.140	.0092	.427	2.168	.0087	.375	2.232	.0073	.268
	Media ítem	2.141	.0092	.425	2.170	.0086	.372	2.245	.0072	.261
	Media sujeto	2.124	.0099	.487	2.140	.0099	.487	2.179	.0098	.476
	Media suj. corregida	2.125	.0098	.484	2.142	.0098	.482	2.191	.0096	.464
	Regresión	2.128	.0097	.469	2.147	.0095	.453	2.202	.0088	.388
	EM	2.128	.0097	.470	2.145	.0096	.456	2.193	.0090	.406
	EM con var. auxiliar	2.124	.0099	.487	2.140	.0099	.487	2.179	.0097	.474
Condicionada NMAR 80/20	Listwise	2.248	.0102	.407	2.371	.0099	.314	2.507	.0100	.226
	Media escala	2.158	.0091	.416	2.206	.0083	.344	2.287	.0071	.254
	Media ítem	2.159	.0091	.413	2.212	.0082	.337	2.297	.0070	.248
	Media sujeto	2.134	.0098	.481	2.157	.0097	.471	2.218	.0095	.448
	Media suj. corregida	2.135	.0098	.478	2.163	.0096	.460	2.228	.0094	.438
	Regresión	2.139	.0096	.462	2.175	.0092	.419	2.244	.0085	.362
	EM	2.139	.0096	.463	2.171	.0092	.427	2.232	.0088	.385
	EM con var. auxiliar	2.135	.0098	.478	2.168	.0095	.452	2.231	.0093	.434

Por su parte, los métodos de sustitución por la regresión y el algoritmo EM presentan una media similar a los mejores métodos, aunque acusan una mayor infraestimación del error típico y de la varianza poblacional. Se considera que estos dos procedimientos se encuentran a medio camino entre los mejores y los peores métodos de imputación.

En todo caso es necesario advertir que cualquier método de imputación tiene sus limitaciones cuando las condiciones de pérdida se vuelven muy adversas. Así en la condición NMAR 80/20 y 20% el método de imputación con mejor funcionamiento (reemplazo por la media del sujeto) sobreestima la media original un 5%.

Los resultados sobre dimensionalidad y consistencia interna de la escala se presentan en la tabla 6. Las conclusiones son similares a las vistas con respecto a los estadísticos básicos de posición y variabilidad. En general cualquier método funciona mejor cuando el mecanismo de pérdida es aleatorio y se deteriora más a medida que la pérdida es mayor y más sesgada hacia los sujetos de bajo autoconcepto académico.

Bajo el mecanismo de pérdida MCAR el método *listwise deletion* es el que mejor se ajusta a los datos originales con independencia del tamaño de la pérdida. Sin embargo, con pérdidas NMAR, este método es desaconsejable. A pesar de esto, lo cierto es que en aquellas ocasiones en que sea posible confirmar una pérdida MCAR los resultados del método *listwise deletion* pueden ser buenos estimadores de la consistencia interna y la dimensionalidad de la escala en los análisis exploratorios iniciales.

Por su parte, los datos de consistencia y dimensionalidad vuelven a mostrar las limitaciones del reemplazo por la media de la escala o por la media del ítem. Ambos procedimientos presentan los peores resultados en las nueve condiciones de pérdida, observándose en todas ellas una bajada de alfa y pérdida explicativa del primer factor.

Los cinco procedimientos restantes presentan un comportamiento similar: ganancia de alfa y aumento del porcentaje de varianza explicado por el primer factor a medida que la pérdida se hace mayor con independencia del mecanismo de pérdida. En todo

caso, este incremento afecta sobremanera a los métodos basados en la información del sujeto (reemplazos por la media y EM con variables auxiliares). Los métodos de reemplazo por la regresión y el algoritmo EM son los que presentan una menor inflación tanto de alfa como del porcentaje de varianza explicada por el primer factor. Ambos presentan valores similares y en pérdidas intencionales son los que mejor ajustan al parámetro original.

Discusión y conclusiones

Los estudios de evaluación de sistemas educativos basados en grandes muestras suelen presentar valores perdidos. La primera conclusión de este trabajo indica, al igual que Botella (2002), que ignorar estos valores perdidos como hace el procedimiento *listwise deletion* no parece una estrategia adecuada. Por tanto, es necesario emplear algún procedimiento de imputación que permita recuperar los datos.

Para elegir el procedimiento de imputación adecuado es necesario explorar el mecanismo de pérdida de los datos y estimar el porcentaje de pérdida de los mismos. Si es posible confirmar que los datos se pierden aleatoriamente y que dichas pérdidas son pequeñas (en torno al 5% de los datos) las diferencias entre los diferentes métodos para recuperar los datos perdidos son pequeñas. En este caso puede estar justificado el uso de los métodos más sencillos, es decir, aquellos con menos complejidad y menos tiempo de cálculo.

Sin embargo, como muestran los datos empíricos presentados al principio del artículo, la pérdida de datos en las evaluaciones diagnósticas de los sistemas educativos no es aleatoria (NMAR), en cuyo caso no todos los métodos de recuperación de datos parecen ser igualmente válidos. Nuestros resultados aconsejan el empleo de una combinación de métodos de imputación en función de las características de los datos perdidos. Si se dispone de un dato incompleto, es decir, cuando se tiene información parcial del su-

Tabla 6
Coeficiente alfa de Cronbach y porcentaje de varianza explicada por el primer factor en la escala autoconcepto en función del mecanismo de pérdida, el tamaño y el método de recuperación de los datos

Mecanismo de pérdida	Método de imputación	Tamaño de la pérdida					
		5%		10%		20%	
		Alfa	% Varianza 1º factor	Alfa	% Varianza 1º factor	Alfa	% Varianza 1º factor
Datos completos (N = 5000)		**.850**	**62.7**	**.850**	**62.7**	**.850**	**62.7**
Aleatoria MCAR	Listwise	.847	62.2	.850	62.6	.850	62.7
	Media escala	.834	60.3	.821	58.4	.785	53.8
	Media ítem	.834	60.3	.822	58.5	.787	54.1
	Media sujeto	.859	64.1	.869	65.8	.890	69.6
	Media suj. corregida	.859	64.1	.870	66.0	.892	69.9
	Regresión	.856	63.6	.864	65.0	.879	67.6
	EM	.856	63.6	.864	65.0	.879	67.6
	EM con var. auxiliar	.859	64.1	.869	65.8	.891	69.6
Condicionada NMAR 60/40	Listwise	.838	61.0	.827	59.6	.778	53.8
	Media escala	.828	59.4	.807	56.6	.747	50.0
	Media ítem	.826	59.2	.804	56.3	.739	49.1
	Media sujeto	.860	64.3	.872	66.4	.893	70.3
	Media suj. corregida	.860	64.3	.872	66.4	.893	70.2
	Regresión	.855	63.5	.863	64.8	.867	65.6
	EM	.856	63.6	.864	65.0	.874	66.9
	EM con var. auxiliar	.860	64.3	.873	66.4	.894	70.4
Condicionada NMAR 80/20	Listwise	.832	60.1	.798	56.1	.764	52.1
	Media escala	.824	58.9	.792	54.9	.743	49.5
	Media ítem	.822	58.6	.784	53.9	.734	48.6
	Media sujeto	.861	64.4	.870	66.0	.889	69.4
	Media suj. corregida	.861	64.4	.869	65.9	.888	69.1
	Regresión	.855	63.5	.852	63.2	.859	64.1
	EM	.855	63.6	.856	63.9	.870	65.9
	EM con var. auxiliar	.861	64.4	.865	65.3	.886	68.9

jeto, el reemplazo del dato perdido por la media del sujeto sin corrección es un procedimiento que aúna precisión para recuperar los parámetros originales con sencillez y rapidez de cálculo. Sin embargo, el reemplazo por la media del sujeto solo es posible cuando el mismo ha respondido a algún ítem de la escala. Si no se dispone de respuesta alguna el caso quedaría definitivamente perdido. Por ello, cuando el caso está completamente vacío puede emplearse el algoritmo EM con variables auxiliares. De entre todas las variables auxiliares disponibles deben elegirse aquellas que presenten una buena correlación con la variable a imputar. Procediendo de este modo, es decir, reemplazando por la media del sujeto los casos incompletos y, en un segundo lugar, aplicando un procedimiento de imputación múltiple como el EM con variables auxiliares para datos completamente perdidos se pueden lograr estimaciones precisas de los estadísticos fundamentales de la escala y también predicciones ajustadas del rendimiento académico. Naturalmente esto tiene sentido a la hora de generar un diagnóstico del sistema educativo, pero plantea serios problemas deontológicos como sistema de asignación de puntuaciones individuales.

Una última consideración debe hacerse con respecto a la sobreestimación de la consistencia interna y del porcentaje de varianza explicada por el primer factor cuando se emplean los métodos de reemplazo por la media del sujeto y el algoritmo EM. Como ya se advirtió la mejora aparente de la fiabilidad de la escala cuando se emplean estos métodos se debe a que ambos incrementan de forma artificial la unidimensionalidad de la escala. Si los datos fuesen perdidos aleatoriamente el método *listwise deletion* parece ofrecer estimaciones ajustadas de estos estadísticos. En cambio, para pérdidas intencionales se recomienda utilizar como estimadores de fiabilidad y dimensionalidad los resultados de los métodos de regresión o el algoritmo EM sin variables auxiliares.

Esto resulta esperable debido a la propia naturaleza de la operación que se realiza, pues al cambiar el dato perdido por información del propio sujeto lo que se está haciendo es acentuar las diferencias entre los sujetos e inyectar unidimensionalidad a la escala de forma artificial, si bien desde un punto de vista psicológico tiene mucho sentido que lo que mejor prediga la actuación global del individuo sea su ejecución parcial.

La generalización de estos resultados debe de hacerse de forma cautelosa, nótese que están obtenidos a partir de una escala de Autoconcepto de cinco ítems claramente unidimensional. Bajo estas condiciones es esperable que los métodos de imputación basados en el sujeto funcionen razonablemente bien, dado que todos los ítems convergen en una dimensión. Sin embargo, cabría preguntarse si su eficacia se mantiene cuando se trabaja con constructos con menor consistencia interna (Botella y Ponte, 2011), o incluso cuando se trabaja con variables con dos o más dimensiones claras. Sería necesario saber si bajo condiciones menos favorables en cuanto a fiabilidad y claridad del constructo la parte, esto es, la respuesta a un ítem, sigue siendo el mejor predictor del todo, esto es, la puntuación en la escala.

De igual modo la variable autoconcepto tiene unas características bien definidas: es un constructo psicológico, fuertemente conectado al criterio empleado para comprobar su capacidad de predicción (la competencia matemática en este caso) y que ha sido tratada como una variable cuantitativa, promediando las respuestas de los sujetos a ítems con cuatro categorías de respuesta. Faltaría saber si los métodos de reemplazo por la media del sujeto funcionan igual de bien cuando se trabaja con variables categóricas (García-Fernández, Secades-Villa, García-Rodríguez, Álvarez-López, Sánchez-Hervás, Fernández-Hermida y Fernández-Artamendi, 2011). De igual modo cabría preguntarse si este método de imputación sigue funcionando en variables de carácter sociológico como las profesiones y estudios de los padres, y en cuya imputación habría que tener en cuenta el contexto general del centro en el que se escolariza el alumnado.

Referencias

Bokossa, M.C., y Huang, G.G. (2001). *Imputation of tests scores in the National Education Longitudinal Study of 1988 (NELS: 88)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Botella, J. (2002). Potencia de pruebas alternativas para dos muestras relacionadas con datos perdidos. *Psicothema, 14*(1), 174-180.

Botella, J., y Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23*, 516-522.

Cohen, J., y Cohen, P. (1985). *Applied multiple regresion and correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S.G., y Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioural sciences* (3ª ed.). Mahwah, N.J.: Lawrence Erlbaum.

Downey, R.G., y King, C.V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *The Journal of General Psychology, 125*(2), 175-191.

Enders, C.K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*(3), 419-436.

Fernández, R., y Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula Abierta, 39*(2), 3-34.

García-Fernández, G., Secades-Villa, R., García-Rodríguez, O., Álvarez-López, E., Sánchez-Hervás, E., Fernández-Hermida, J.R., y Fernández-Artamendi, S. (2011). Individual characteristics and response to contingency management treatment for cocaine addiction. *Psicothema, 23*, 114-118.

Gobierno del Principado de Asturias (2008). Evaluación de Diagnóstico Asturias 2008. Oviedo, Consejería de Educación y Ciencia. Consultado el 17 de mayo de 2011 en: http://www.educastur.es/media/publicaciones/informes/evadiag2008b.pdf.

Gobierno del Principado de Asturias (2010). Evaluación de Diagnóstico Asturias 2009. 2º curso de Educación Secundaria Obligatoria, Oviedo, Consejería de Educación y Ciencia. Consultado el 17 de mayo de 2011 en: http://www.educastur.es/media/institucional/calidad/diagnostico_ast09/ED_2009.pdf.

Gobierno del Principado de Asturias (2011). Evaluación de diagnóstico Asturias 2010. 4º curso de Educación Primaria, Oviedo, Consejería de Educación y Ciencia. Consultado el 18 de julio de 2011 en: http://www.educastur.es/media/publicaciones/informes/evadiag2010.pdf.

Graham, J.W. (2009). Missing data analysis: Making it work in the real World. *Annual Review of Psychology, 60*, 549-576.

Howell, D.C. (2007). The analysis of missing data. En W. Outhwaite y S. Turner (Eds.), *Handbook of social science methodology*. London: Sage.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198-1202.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87*(420), 1227-1237.

Little, R.J.A., y Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Ministerio de Educación (2009). *Evaluación General de Diagnóstico 2009. Marco de la Evaluación*. Madrid: Instituto de Evaluación.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Trong, K.L., y Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C., y Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Centre, Boston College.

National Assessment Governing Board (2003). *Background information framework for the National Assessment of Educational Progress*. Washington, DC: NAGB, U.S. Department of Education.

OECD (2009). PISA 2009 assessment framework. Key competencies in reading, mathematics and science. París: OECD.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*, 3-15.

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

San Luis, C., Hernández, J.A., y Ramírez, G. (1997). Estimación de datos perdidos por máxima verosimilitud en patrones «missing» aleatorios (MAR) y completamente aleatorios (MCAR) en modelos estructurales. *Psicothema, 9*(1), 187-197.

Shrive, F.M., Stuart, H., Quan, H., y Ghali, W.A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology, 6,* 57. Consultado el 16 de septiembre en: http://www.biomedcentral.com/1471-2288/6/57.

Van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology*, *6*(1), 31-36.

Van Ginkel, J.R., Sijtsma, K., Van der Ark, L.A., y Vermunt, J.K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*, *6*(1), 17-30.

Van Ginkel, J.R., y Van der Ark, L.A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, *29*(2), 152-153.

Von Hippel, P.T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician, 58*(2), 160-164.

Vansteelandt, S., Carpenter, J., y Kenwark, M.G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology, 6*(1), 37-48.

Willms, J.D., y Smith, T. (2006). *A Manual for Conducting Analyses with Data from TIMSS and PISA (Report prepared for the UNESCO Institute for Statistics)*. New Brunswick: Canadian Research Institute for Social Policy. Consultado el 17 de mayo de 2011 en: http://www.unb.ca/crisp/pdf/Manual_TIMSS_PISA2005_0503.pdf.

Yamamoto, K. (2001). Estimating literacy proficiencies with and without cognitive data. En National Center for Education Statistics: *Technical report and data file user's manual for the 1992 National Adult Literacy Survey*. Washington, DC: U.S. Department of Education (pp. 142-164).