Psicothema

Psicothema 2014, Vol. 26, No. 4, 516-523
doi: 10.7334/psicothema2014.98

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Estimating the reliability coefficient of tests in presence of missing values

Marcelino Cuesta Izquierdo[1] and Eduardo Fonseca Pedrero[2]
[1] Universidad de Oviedo and [2] Universidad de La Rioja

Abstract

Background: The problem of missing values at the item level is common in studies using educational and psychological tests. The aim of the present work is to explore how the estimation of reliability is affected by missing values. **Method:** Using real data, we simulated missing values in accordance with a "missing at random mechanism". Four factors were manipulated with the aim of checking their effect on the estimation of the reliability of the instrument: missing data mechanism, percentage of missing data in the database, sample size, and procedure employed for the treatment of missing values. **Results:** The results show that the quality of estimations depends on the interaction of various factors. The general tendency is that the estimations are worse when the sample size is small and the percentage of missing values increases. Listwise is the worst procedure for treatment of the missing data in the simulated conditions. **Conclusions:** It is concluded that with a small percentage of missing values one can obtain estimations that are acceptable from a practical point of view with all the procedures employed, except Listwise.

Keywords: missing values, reliability, imputation, missing at random, test.

Resumen

Estimación del coeficiente de fiabilidad en presencia de valores perdidos. **Antecedentes:** el problema de la presencia de valores perdidos a nivel de ítem es común en los estudios que emplean tests psicológicos o educativos. El objetivo de este trabajo es explorar cómo se ve afectada la estimación de la fiabilidad por la presencia de valores perdidos. **Método:** partiendo de datos reales se simularon valores perdidos de acuerdo a un mecanismo aleatorio. Se manipularon cuatro factores para comprobar su influencia en la estimación de la fiabilidad de la prueba: mecanismo de pérdida de datos, porcentaje de valores perdidos, tamaño de muestra y método empleado para el manejo de los datos perdidos. **Resultados:** los resultados muestran que la calidad de las estimaciones depende de la interacción de varios factores. La tendencia general es que las estimaciones son peores cuando el tamaño de muestra es pequeño y aumenta el porcentaje de valores perdidos. Listwise es el peor procedimiento de manejo de los valores perdidos en las condiciones simuladas. **Conclusiones:** cuando el porcentaje de valores perdidos es pequeño pueden obtenerse estimaciones aceptables, desde un punto de vista práctico, con todos los procedimientos empleados, excepto Listwise.

Palabras clave: valores perdidos, fiabilidad, imputación, valores perdidos aleatorios, test.

In the Health, Behaviour, and Educational Sciences, researchers often make use of tests and questionnaires to obtain data. When such instruments are applied, some participants commonly fail to answer one or more items. As van der Ark and Vermunt (2010) point out, ignoring the problem can lead to statistically biased results and erroneous conclusions.

Researchers have been concerned about the problem of missing values for a long time, but it was not until the end of the last century that it began to be studied systematically (Graham, 2009; Little & Rubin, 1987; Rubin, 1987). Currently, the missing data mechanisms defined by Rubin (1976) are well established in the literature: (a) data missing completely at random (MCAR); (b) data missing at random (MAR); and (c) missing not at random (MNAR).

But, as Howell (2008) points out, despite the fact that the treatment of these missing values is not an especially controversial matter at the statistical level, there does not appear to be a good flow of such knowledge from the statistical-methodological context to applied fields (Baraldi & Enders, 2010; Graham, 2009; Roth, 1994; Schafer & Graham, 2002).

The various procedures proposed for the treatment of missing values can be grouped into the so-called traditional methods and modern methods. A distinction is made in the traditional method category between: (a) deletion methods, which would include some highly popular procedures such as *Listwise* (analysis of complete cases) or *Pairwise* (analysis of available cases); and (b) simple imputation methods, such as using some type of mean (of the scale, of the item, of the respondent, etc.), deterministic or stochastic regression, or the *Hot Deck* procedures. Modern methods would include maximum-likelihood and multiple imputation procedures.

Although in the statistical literature, the superiority of the so-called modern procedures appears to be well-established for the case of MAR missing data mechanisms (Allison, 2002; Enders, 2010), in more applied contexts, it is still customary for researchers to use simpler procedures. For example, in a study on personality tests, Van Ginkel, Sijtsma, van der Ark and Vermunt (2010) found that the most widely used procedure was *Listwise*. As Sijtsma and

Received: April 25, 2014 • Accepted: August 6, 2014
Corresponding author: Marcelino Cuesta Izquierdo
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: mcuesta@uniovi.es

van der Ark (2003) note, this may be due to the fact that applied researchers tend to opt for procedures that are simpler to apply or can be implemented via standard statistical programs.

When data-collection procedures involve tests or questionnaires, a key aspect to take into account are the psychometric properties of such instruments. As McDonald, Thurstone and Nelson (2000) stress, some psychometric properties, such as the reliability or the variance of the scale, influence the covariation between the variable measured by the test and other variables. If, in situations of missing values, the procedure employed for dealing with them introduces any type of bias, this will have effects on the relations between the variable measured and other variables, which will influence some of the procedures to gather validity evidence (Ríos & Wells, 2014; Oren, Kennet-Cohen, Turvall, & Allalouf, 2014).

As Roth, Switzer and Switzer (1999) pointed out, up until then, there was scarcely any literature on the problem of missing values at the item level. Since then, there have been numerous studies in this field, though not always dealing directly with psychometric aspects (Bernaards & Sijtsma, 2005; Carpita & Manisera, 2011; Cuesta, Fonseca-Pedrero, Vallejo, & Muñiz, 2013; Enders, 2003, 2004; Fernández-Alonso, Suárez-Alvarez, & Muñiz, 2012; Gmel, 2001; McDonald, Thurston, & Nelson, 2000; Shrive, Stuart, Quan, & Ghali, 2006; Sijtsma & van der Ark, 2003; Van Ginkel, van der Ark, & Sijtsma, 2007a, 2007b).

The aim of the present work is to contribute to the accumulated information on the way estimation of the reliability coefficients of tests is influenced by missing values that follow a MAR mechanism. We used coefficient alpha, as this is the reliability coefficient reported more commonly in empirical studies (López-Pina, Sánchez-Meca, & López-López, 2012). Given the applied nature of this study, we shall be working from Raaijmakers' (1999) perspective, according to which the choice of a particular procedure for the treatment of missing values will depend on two factors: first, it is necessary to take into account the procedures available to applied researchers and the software that permits their easy application; second, the effectiveness of the different procedures in a certain research context, given a range of conditioning factors such as sample size, percentage of missing values, missing value mechanism, and so on. We chose to test here the methods implemented in a software package widely used in applied research, namely SPSS.

Method

Participants

Participants were the same as those who took part in the construction of the Oviedo Questionnaire for the Assessment of Schizotypy (ESQUIZO-Q) (Fonseca-Pedrero, Muñiz, Lemos, Paíno, & Villazón, 2010). The sample was obtained by means of random stratified cluster sampling, at the classroom level, and in the Spanish Autonomous Region of the Principality of Asturias. The final sample size was of 3,056 youngsters, of whom 48.1% were boys. Age range was 14 to 18 years, with a mean of 15.9 years and a standard deviation of 1.17.

Instrument

The ESQUIZO-Q (Fonseca-Pedrero et al., 2010) is a self-report designed to assess schizotypal personality traits in adolescent population. The instrument is made up of 51 items with Likert-type response format and 5 categories.

The psychometric properties of the ESQUIZO-Q have been widely analyzed from Classical Test Theory (Fonseca-Pedrero, Lemos-Giráldez et al., 2011; Fonseca-Pedrero, Paíno et al., 2011).

Design

Four factors were manipulated with the aim of checking their effect on the estimation of the instrument's reliability measured by means of Cronbach's alpha coefficient: missing data mechanism, percentage of missing data in the database, sample size, and procedure employed for the treatment of missing values.

Missing data mechanism: Data were missing according to a MAR mechanism as a function of the sex variable. Two situations were considered. In the first of these, the probability of missing values in the girls' group was double that of the boys' group (MAR 2 to 1); in the second case, the probability in the girls' group was three times that of the boys' group (MAR 3 to 1).

Percentage of missing values: 5%, 10%, 20% and 30% were considered.

Sample sizes: 3056 cases (total sample) and 200 (random subsamples taken from the total sample).

Procedure for the treatment of missing values: Five procedures were used:

Listwise: Eliminating from the analyses those participants with missing values in any of the variables to be analyzed.

Imputation by means of *multiple linear regression*. Item score is imputed by means of a multiple regression model using the scores of participants with all the responses, with the missing-value item as dependent variable and the rest of the items as independent variables. Added to the score predicted from the model is a random error extracted from a normal distribution (with mean 0 and standard deviation equal to the square root of the mean squared error term of the regression).

Imputation by means of the *EM* (*expectation-maximization*) procedure. This is an algorithm that permits estimations of maximum-likelihood through a two-step procedure. In the first step (E), values are generally imputed using regression equations, and in the second step (M), the values are calculated again for the means and the covariance matrix using the imputed values and not the missing values. Once the new estimations of the means and covariances have been obtained, the process begins again with step E, and continues until the estimations converge. Resulting from the imputation are variance-covariance matrices from which the Cronbach's alpha coefficient is subsequently calculated.

Imputation in the final imputation cycle by means of EM, which we have called "*Simple EM imputation*". The SPSS program offers the possibility of imputing the raw data after the final cycle of EM, which is actually an imputation via regression using a maximum-likelihood estimation of the vector of means and of the variance-covariance matrix (Enders, 2010).

Multiple Imputation. SPSS uses a sequential regression procedure (*fully conditional specification*). This is an interactive model, in which for each interaction and for each variable employed, the method fits a model with one dependent variable and uses all the rest on the list as predictors, so that the missing values for the variable being fitted can be imputed. The procedure continues until the specified maximum number of interactions is reached and the values imputed in the final interaction are saved in the imputed database. We used the SPSS default options, with 5 imputations and 10 interactions. The estimations of Cronbach's alpha calculated for each of the five databases imputed are averaged by means of the formulas proposed by Rubin (1987), so that a single final estimation of the coefficient is obtained.

Simulation procedure

The starting point for the generation of data for the present study was the complete matrix of responses of the 3,056 participants to the 51 items. Based on this matrix, we generated MAR models for each item, obtaining the percentage of missing values previously established, respecting the different probabilities of generating a missing value depending on whether the participant was male or female.

To obtain the desired missing values mechanism for each item, we generated, using SPSS, a random variable with uniform distribution for the boys and another one for the girls. When the random variable yielded values lower than the proportion of missing values we wished to obtain, the item value was converted into a missing value, without setting limits for the number of missing values a participant could present. For each one of the missing values treatment procedures, 100 databases were generated.

To generate the 200-size samples, we introduced a previous step that consisted in generating, using SPSS, random subsamples without replacement based on the total database, to which we subsequently applied the procedure described above.

Data analysis

On the basis of the original matrix we calculated the Cronbach's α coefficient that was taken as a reference for the values obtained in the matrices generated.

For each matrix we calculated the value of α, and based on its values, two indicators of the differences between the value of the original matrix and the estimations in the imputed data.

Root mean square error (RMSE), which is the average of the difference between $\hat{\alpha}$ (the reliability estimated in the imputed data) and α (the reliability of the original complete matrix), and which is used as an indicator of the variability of the estimations

$$RMSE = \sqrt{\frac{\Sigma(\hat{\alpha}-\alpha)^2}{100}}$$

Also calculated was the *average bias*, following the expression

$$Bias = \frac{\Sigma(\hat{\alpha}-\alpha)}{100}$$

For the sample size 200 we also obtained the difference between the Cronbach's alpha value calculated for a complete

Table 1
Descriptive data for the databases with n = 3056

	% of missing values proposed	% of values missing			% of cases affected			number of items missing in the cases affected		
		min	max	mean	min	max	mean	min	max	mean
MAR 2 to 1	5	3.7	6.6	5.07	88.9	91.5	89.9	1	16	2.88
	10	8.3	12.1	10.14	97.9	99.1	98.6	1	19	5.25
	20	17.8	22.7	20.25	99.86	100	99.97	1	29	10.33
	30	27.6	33.4	30.83	99.97	100	99.99	3	37	15.5
MAR 3 to 1	5	3.7	6.5	5.1	84.5	87.3	85.8	1	16	3.03
	10	8.1	12.4	10.19	95.7	97.4	96.5	1	22	5.39
	20	17.8	22.8	20.39	99.5	99.9	99.8	1	31	9.81
	30	27.9	33.4	30.57	99.93	100	99.99	1	39	15.59

Table 2
Descriptive data for the databases with n = 200

	% of missing values proposed	% of values missing			% of cases affected			number of items missing in the cases affected		
		min	max	mean	min	max	mean	min	max	mean
MAR 2 to 1	5	0.5	11.5	5.07	85	96	89.8	1	11	2.88
	10	3.5	19.5	10.13	96	100	98.5	1	18	5.25
	20	10.5	30.5	20.25	99	100	99.98	1	27	10.33
	30	20.5	42	30.47	100	100	100	1	36	15.54
MAR 3 to 1	5	1	12	5.09	78.5	92.5	85.6	1	13	3.03
	10	2.5	18	10.23	92.5	99.5	96.6	1	18	5.34
	20	10.5	31.5	20.38	99	100	99.8	1	30	10.42
	30	18	42	30.43	99.5	100	99.9	1	36	15.52

200-size base ($\hat{\alpha}_{c_i}$) and the value calculated for a database in which some procedure for the treatment of missing values had been applied, and which had been generated on the basis of the above-mentioned complete database ($\hat{\alpha}_{I_i}$). We called the resulting value the *discrepancy* ($|\hat{\alpha}_{c_i} - \hat{\alpha}_{I_i}|$) (Van Ginkel, van der Ark, & Sijtsma, 2007), and on the basis of this value we obtained the *average discrepancy*.

$$Avg.Disc = \frac{\Sigma\left|\hat{\alpha}_{C_i} - \hat{\alpha}_{I_i}\right|}{100}$$

Results

Descriptive results

The general tendency is for the results to be poorer in the 200-size samples.

RMSE

In the 3056-size samples the variability increases as the percentage of missing values increases, but in smaller samples, there is no uniform pattern of behaviour: for example, *Simple EM* and *EM* yield lower values with 20% of missing values than with 5%. As regards the missing values mechanism, in large samples, there are no differences between MAR 2 to 1 and MAR 3 to 1; once again, the results are less clear in the small samples, given that for some procedures, there is a slight improvement in the case of MAR 2 to 1, and in others, in that of MAR 3 to 1.

In the comparison of procedures for the treatment of missing values, the *Listwise* method emerges as that which yields the poorest results, regardless of sample size, *EM* is that which yields better results in large samples, and, surprisingly, *Simple EM* the one that yields better results in small samples.

Bias

In general, it was found that all the methods tend towards underestimation except *simple EM*, which overestimates. Likewise, we observed an increase in average bias as the percentage of missing values increased. As regards the missing value mechanism, in the large sample sizes, performance becomes slightly poorer on passing from MAR 2 to 1 to MAR 3 to 1; however, in the case of small samples, there is no single pattern for the different methods.

As occurred with regard to variability, the method with the poorest behaviour is *Listwise*, while the best-behaved are *EM* for large samples and *simple EM* for small samples.

Discrepancy

As in the cases of the two previous indicators, behaviour becomes poorer as the percentage of missing values increases, with no notable differences between MAR 2 to 1 and MAR 3 to 1. Once again, *Listwise* shows the poorest behaviour, and *EM* yields the best results.

It should be noted that in the 200-size samples with 30% of missing values, the regression procedure values show a marked increased in all three indexes considered here, in comparison with both its own behaviour in the case of a lower percentage of missing values, and with the other procedures with this same rate of missing values (see Table 3).

Inferential results

In addition to the descriptive analyses reported in the previous section, ANOVAS were used for identifying which of the factors manipulated were related in statistically significant fashion to the results obtained in the three dependent variables (RMSE, bias, discrepancy). Three analyses were carried out. In the first two (Tables 5 and 6) we took as dependent variables the mean values

Table 3
Mean values of RMSE and bias for the different missing value treatment procedures and percentages of missing values (n = 3056)

α = .8857		**MAR 2 to 1**					**MAR 3 to 1**				
		Procedure					**Procedure**				
% of missing		**Listwise**	**Regression**	**Simple EM**	**Cov EM**	**MI**	**Listwise**	**Regression**	**Simple EM**	**Cov EM**	**MI**
	Mean α	.8856	.8854	.8896	.8856	.8855	.8866	.8853	.8896	.8856	.8854
5	RMSE	.0106	.0006	.0039	.0003	.0004	.0082	.0006	.0039	.0004	.0004
	Mean bias	-.0001	-.0004	.0039	-.00007	-.0002	.0009	-.0004	.0039	-.0001	-.0003
	Mean α	.8775	.8849	.8936	.8855	.8852	.8832	.8850	.8936	.8856	.8853
10	RMSE	.0376	.0010	.0079	.0005	.0007	.0226	.0009	.0079	.0005	.0006
	Mean bias	-.0082	-.0008	.0078	-.0002	-.0005	-.0025	-.0007	.0079	-.0001	-.0004
	Mean α		.8840	.9016	.8854	.8847		.8836	.9015	.8852	.8844
20	RMSE		.0020	.0159	.0008	.0012		.0022	.0158	.0008	.0015
	Mean bias		-.0018	.0159	-.0003	-.0010		-.0021	.0158	-.0005	-.0013
	Mean α		.8818	.9097	.8850	.8838		.8816	.9100	.8851	.8838
30	RMSE		.0042	.0240	.0012	.0022		.0043	.0243	.0011	.0021
	Mean bias		-.0039	.0240	-.0007	-.0019		-.0041	.0243	-.0006	-.0019

Simple EM: imputation of the values of the final cycle of the Expectation-Maximization procedure; **Cov EM**: imputation by Expectation-Maximization; **MI**: multiple imputation

Table 4
Mean values of RMSE, bias and discrepancy for the different missing value treatment procedures and percentages of missing values. (n=200)

α = .8857		**MAR 2 to 1**						**MAR 3 to 1**					
		Procedure						**Procedure**					
% of missing		**Comp**	**Listwise**	**Regression**	**Simple EM**	**Cov EM**	**MI**	**Comp**	**Listwise**	**Regression**	**Simple EM**	**Cov EM**	**MI**
5	Mean α	.8838	.8676	.8817	.8857	.8829	.8803	.8844	.8732	.8828	.8867	.8825	.8814
	RMSE	.0148	.0551	.0156	.0145	.0150	.0160	.0140	.0393	.0148	.0141	.0154	.0151
	Mean bias	-.0019	-.0181	-.0040	.00002	-.0028	-.0054	-.0014	-.0125	-.0029	.0010	-.0032	-.0043
	Mean discrepancy		.0162	.0021	-.0019	.0009	.0034		.0112	.0016	-.0024	.0018	.0029
10	Mean α	.8844		.8803	.8880	.8827	.8774	.8831	.8142*	.8799	.8871	.8817	.8762
	RMSE	.0126		.0144	.0128	.0135	.0160	.0149	.0199	.0165	.0147	.0158	.0186
	Mean bias	-.0014		-.0054	.0023	-.0030	-.0083	-.0026	-.0740	-.0058	.0014	-.0040	-.0095
	Mean discrepancy			.0040	-.0036	.0016	.0069		.0709	.0032	-.0040	.0014	.0069
20	Mean α	.8858		.8573	.8899	.8809	.8704	.8850		.8628	.8898	.8802	.8689
	RMSE	.0120		.0332	.0131	.0142	.0212	.0131		.0278	.0133	.0147	.0222
	Mean bias	.0001		-.0284	.0042	-.0048	-.0154	-.0007		-.0230	.0041	-.0055	-.0168
	Mean discrepancy			.0285	-.0040	.0049	.0155			.0222	-.0048	.0049	.0161
30	Mean α	.8833		.7562	.8882	.8750	.8528	.8838		.7948	.8907	.8764	.8541
	RMSE	.0121		.1338	.0126	.0172	.0366	.0148		.0967	.0155	.0187	.0371
	Mean bias	-.0024		-.1295	.0025	-.0107	-.0329	-.0019		-.0909	.0050	-.0093	-.0316
	Mean discrepancy			.1272	-.0049	.0083	.0306			.0890	-.0068	.0074	.0297

* 96 databases
Comp: Data set without missing values; **Simple EM**: imputation of the values of the final cycle of the Expectation-Maximization procedure; **Cov EM**: imputation by Expectation-Maximization; **MI**: multiple imputation

Table 5
ANOVA for RMSE

Factors	**SS**	**df**	**MS**	**F**	**p**	**η^2**
Sample	.006	1	.006	295.161	.000	.970
MAR	1.243E-5	1	1.243E-5	.579	.466	.060
Missings	.004	3	.001	57.278	.000	.950
Proc	.002	3	.001	29.942	.000	.909
MAR * Missings	5.715E-5	3	1.905E-5	.888	.484	.228
MAR * Proc	9.614E-5	3	3.205E-5	1.494	.281	.332
Sample * MAR	1.332E-5	1	1.332E-5	.621	.451	.065
Missings * Proc	.004	9	.000	22.049	.000	.957
Sample * Missings	.002	3	.001	27.612	.000	.902
Sample * Proc	.004	3	.001	54.976	.000	.948
MAR * Missings * Proc	.000	9	2.135E-5	.995	.503	.499
Sample * MAR * Missings	5.872E-5	3	1.957E-5	.912	.473	.233
Sample * MAR * Proc	9.699E-5	3	3.233E-5	1.507	.278	.334
Sample * Missings * Proc	.005	9	.001	26.772	.000	.964
Error	.000	9	2.145E-5			

Sample: Sample size; **MAR**: Missing mechanism; **Missings**: % of missing values; **Proc**: missing data handling procedure

Table 6
ANOVA for Bias

Factors	**SS**	**df**	**MS**	**F**	**p**	**η^2**
Sample	.004	1	.004	215.060	.000	.960
MAR	3.285E-5	1	3.285E-5	1.697	.225	.159
Missings	.003	3	.001	44.227	.000	.936
Proc	.006	3	.002	102.621	.000	.972
MAR * Missings	8.952E-5	3	2.984E-5	1.542	.270	.339
MAR * Proc	9.034E-5	3	3.011E-5	1.556	.267	.342
Sample * MAR	3.409E-5	1	3.409E-5	1.762	.217	.164
Missings * Proc	.006	9	.001	36.590	.000	.973
Sample * Missings	.004	3	.001	63.543	.000	.955
Sample * Proc	.002	3	.001	33.998	.000	.919
MAR * Missings * Proc	.000	9	1.865E-5	.964	.522	.491
Sample * MAR * Missings	8.725E-5	3	2.908E-5	1.503	.279	.334
Sample * MAR * Proc	9.336E-5	3	3.112E-5	1.608	.255	.349
Sample * Missings * Proc	.004	9	.000	23.067	.000	.958
Error	.000	9	1.935E-5			

Sample: Sample size; **MAR**: Missing mechanism; **Missings**: % of missing values; **Proc**: missing data handling procedure of RMSE and the bias, respectively, and as factors, the sample size, the missing data mechanism, the percentage of missing values and the procedure for the treatment of missing values (the Listwise method was excluded because with high percentages of missing values all the participants are eliminated). The third analysis (Table 7) was carried out taking discrepancy as the dependent variable and the same factors as in the previous analysis, except for sample size. In the analyses, all the interactions were included, except that of the highest order, which was excluded so as to be able to estimate the error term necessary for applying the F test.

As can be seen in Tables 5 and 6, the pattern is the same for RMSE and bias, with a statistically significant interaction between sample size, percentage of missing values, and procedure for the treatment of missing values, and a large effect size. The rest of the significant interactions are subsumed in this one; on the other hand, it can be observed that the missing data mechanism does not emerge as an influential factor in any of the cases. For the case of discrepancy, the only significant factor was procedure for the treatment of missing values.

Discussion and conclusions

In general terms, we can state that reliability coefficients estimations will be reasonably good as long as the total percentage of missing responses does not exceed 10%. This is applicable to all procedures used here, except for *Listwise*. This is in line with the findings reported in the literature, according to which a necessary condition for *Listwise* to provide acceptable estimations is that the missing data mechanism is MCAR (Botella, 2002; Howell, 2008; Enders, 2010), and also with the results obtained in a previous study where this procedure worked well in such conditions (Cuesta, Fonseca-Pedrero, Vallejo, & Muñiz, 2013). Estimation of the internal consistency of the instrument deteriorates as the number of missing values increases, though differentially depending on the imputation procedure and the sample size.

As follows from previous findings in the general literature on missing values, the maximum-likelihood and multiple imputation procedures should offer the best results. On the whole, it can be stated that our results meet those expectations, even if it appears that maximum-likelihood offers a slightly closer-to-optimum performance compared to multiple imputation. This behaviour does not correspond to the arguments of Gottschal, West and Enders (2012), according to whom multiple imputation should be more flexible when the imputation is carried out at the item level compared to when it is carried out at the scale level; nevertheless,

Table 7
ANOVA for discrepancy

Factors	SS	df	MS	F	p	η^2
MAR	.003	1	.003	1.478	.255	.141
Missings	.011	3	.004	1.893	.201	.387
Proc	.025	3	.008	4.102	.043	.578
MAR * Missings	.006	3	.002	.992	.440	.248
MAR * Proc	.008	3	.003	1.420	.300	.321
Missings * Proc	.025	9	.003	1.369	.324	.578
Error	.018	9	.002			

MAR: Missing mechanism; **Missings**: % of missing values; **Proc**: missing data handling procedure

in other contexts the results do indeed endorse the findings we obtained here (Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011).

Also worthy of mention are the results obtained with the procedure we have called *Simple EM imputation*, insofar as when we work with a large sample, its results deteriorate – even those obtained on imputing with multiple linear regression, which can be considered a variant –, as well as the surprisingly good results obtained when the sample size is small. There is undoubtedly a need for further research so as to obtain a more comprehensive idea of what can be expected from this procedure. With regard to multiple regression, it should be stressed that, according to our results, it is not strongly recommended for high percentages of missing values and small sample sizes.

Sample size appears to be a factor that future research should explore in more depth. When dealing with a large sample size (n = 3,056), the results are in line with what we expected. However, with a sample of 200 individuals, much more realistic in the context of applied work, the patterns found do not reveal clear lines, and it is perhaps on this aspect that we should focus our efforts, with a view to offering practical recommendations for applied researchers.

Finally, we feel it appropriate to make some observations of a practical nature about the use of SPSS for the handling of missing values. The standard procedure used by the program is the highly popular *Listwise,* whose problems are well documented in the literature. In some ways we could argue that this popular software encourages researchers to employ a procedure that is far from optimum. For the researcher who dares to go beyond the default options, SPSS incorporates a model for the *Analysis of missing values*, and another for *Multiple Imputation*; from these, the user can accede to some traditional procedures such as *Listwise*, *Pairwise* and Regression, and to more modern ones such as Expectation-Maximization and Multiple Imputation.

With regard to these last two procedures allow us to highlight some practical issues that may not be immediately "transparent" for the applied researcher. First, the maximum-likelihood procedures estimate vectors of means and variance-covariance matrices, not individual scores. Second, the fact that SPSS offers, in its EM procedure options, the possibility of imputing the raw data after the final cycle of the procedure may be misleading for users, creating in them the illusion that they are using some type of maximum-likelihood imputation, when what they are actually using is nothing more than another version of a regression procedure with the same limitations (Enders, 2010; von Hippel, 2004). And third, when Multiple Imputation is employed, SPSS only implements the *pooling phase* for some statistical procedures.

Focusing on the "Reliability" procedure employed in the present work, the general issues mentioned above involve a series of aspects to take into account. Given that the procedure for calculating the reliability uses as an input the item scores, and that maximum-likelihood imputation estimates a covariance matrix, intermediate steps are necessary for the calculation of the alpha coefficient, either through SPSS syntax itself or that of other software. If multiple imputation is used, it should be borne in mind that the "Reliability" procedure is not one of those equipped with automatic implementation of the pooling phase, and that users must implement it themselves. In sum, it would appear that when in a context of reliability estimation a user wishes to employ the procedures for handling missing values that in this and many other studies are considered the most appropriate, SPSS, despite its popularity, does not offer such user-friendly tools as might be expected for those researchers lacking expertise in methodological aspects.

From the results obtained we can conclude, then, that the so-called modern procedures offer a better performance in the treatment of situations of missing values at the item level in the context of reliability estimation from the classical tests model, and that SPSS, in spite of its popularity, does not appear to be particularly helpful for the use of these procedures in this context.

Anyway, as Allison (2002) states, the only good solution to the problem of missing data is to avoid them. We strongly recommend applied researchers to be careful in the way they design their studies and collect data in order to minimize the missing data and, if they are present, to avoid non-random mechanisms.

On the other hand, our study also shows certain limitations that suggest new research lines, such as: (a) using a large number of items (in the present study we worked with the reliability of the global scale), so that we could work with subscales with a smaller number of items; (b) considering reliability from the perspective of item response theory and substituting the alpha coefficient by the information function at the item and test levels; and (c) in contrast to the use of two quite extreme sample sizes as in the present work, implementing a greater graduation of such sizes in search of possible “critical” sizes. There are, as it can be seen, many aspects in which we can make progress in the direction of providing applied researchers with information on how to address the handling of missing values in their everyday work.

Acknowledgements

This work was funded by the research projects PSI2011-28638 and PSI2011-23095 from the Spanish Ministry of Science and Innovation.

References

Allison, P.D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

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

Bernaards, C.A., & Sijtsma, K. (2005). Bias in factor loadings from questionnaire data with imputed item scores. *Journal of Statistical Computation and Simulation, 75*(1), 13-23.

Botella, J. (2002). Potencia de pruebas alternativas para dos muestras relacionadas con datos perdidos [Power of alternative tests for two paired samples with missing data]. *Psicothema, 14*(1), 174-180.

Carpita, M., & Manisera, M. (2011). On the imputation of missing data in surveys with Likert-type scales. *Journal of Classification, 28*(1), 93-112.

Cuesta, M., Fonseca-Pedrero, E., Vallejo, G., & Muñiz, J. (2013). Datos perdidos y propiedades psicométricas en los tests de personalidad [Missing data and psychometric properties of personality tests]. *Anales de Psicología, 29*(1), 285-292.

Enders, C.K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*(3), 322-337.

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.

Enders, C.K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Fernández-Alonso, R., Suárez-Alvarez, J., & Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas [Imputation methods for missing data in educational diagnostic evaluation]. *Psicothema, 24*(1), 167-175.

Fonseca-Pedrero, E., Muñiz, J., Lemos, S., Paíno, M., & Villazón, U. (2010). *Esquizo-Q. Cuestionario Oviedo para la evaluación de la esquizotipia [ESQUIZO-Q. Oviedo Schizotypy Assessment Questionnaire]*. Madrid: TEA Ediciones.

Fonseca-Pedrero, E., Lemos-Giráldez, S., Paíno, M., Sierra-Baigrie, S., Santarén-Rosell, M., & Muñiz, J. (2011). Internal structure and reliability of the Oviedo Schizotypy Assessment Questionnaire (ESQUIZO-Q). *International Journal of Clinical and Health Psychology, 11*, 385-402.

Fonseca-Pedrero, E., Paíno, M., Lemos-Girádez, S., Sierra-Baigrie, S., Ordóñez, N., & Muñiz, J. (2011). Early psychopathological features in Spanish adolescents. *Psicothema, 23*, 87-93.

Gmel, G. (2001). Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Statistics in Medicine, 20*, 2369-2381.

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

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*, 1-25.

Howell, D.G. (2008). The analysis of missing data. In Outhwaite, W., & Turner, S. (Eds), *Handbook of Social Science Methodology*. London: Sage.

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

López-Pina, J.A., Sánchez-Meca, J., & López-Lopez, J.A. (2012). Métodos para promediar coeficientes alfa en los estudios de generalización de la fiabilidad [Methods for averaging alpha coefficients in reliability generalization studies]. *Psicothema, 24*(1), 161-166.

Mcdonald, R.A., Thurston, P.W., & Nelson, M.R. (2000). A Monte Carlo study of missing item methods. *Organizational Research Methods, 3*(1), 71-92.

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema, 26*(1), 117-126.

Raaijmakers, Q.A.W. (1999). Effectiveness of different missing data treatments in surveys with Likert-Type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59*(5), 725-748.

Ríos, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Roth, P.L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*, 537-560.

Roth, P.L., Switzer, F.S., & Switzer, D.M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods, 2*(3), 211-232.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

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

Schafer, J.J., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Shrive, F., Stuart, H., Quan, H., & 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 (http://www.biomedcentral.com/1471-2288-6-57) (download on 25/01/2011).

Sijtsma, K., & van der Ark, L.A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*(4), 505-528.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero-Herrero, E. (2011). Comparison of modern methods for analyzing unbalanced repeated measures data. *Multivariate Behavioral Research, 46*(6), 900-937.
Van der Ark, L.A., & Vermunt, J.K. (2010). New developments in missing data analysis (editorial). *Methodology, 6*(1), 1-2.
Van Ginkel, J.R., Sijtsma, K., van der Ark, L.A., & 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., van der Ark, L.A., & Sijtsma, K. (2007a). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*(2), 387-414.
Van Ginkel, J.R., van der Ark, L.A., & Sijtsma, K. (2007b). Multiple imputation of item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60*(2), 315-337.
Von Hippel, P.T. (2004). Biases in SPSS 12.0 missing value analysis. *The American Statistician, 58*(2), 160-164.