Psicothema

Psicothema 2017, Vol. 29, No. 3, 384-389
doi: 10.7334/psicothema2016.75

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Comparison of methods for dealing with missing values in the EPV-R

David Paniagua[1], Pedro J. Amor[2], Enrique Echeburúa[3] and Francisco J. Abad[1]
[1] Universidad Autónoma de Madrid, [2] Universidad Nacional de Educación a Distancia and [3] Universidad del País Vasco

Abstract

Background: The development of an effective instrument to assess the risk of partner violence is a topic of great social relevance. This study evaluates the scale of "Predicción del Riesgo de Violencia Grave Contra la Pareja" –Revisada– (EPV-R - Severe Intimate Partner Violence Risk Prediction Scale-Revised), a tool developed in Spain, which is facing the problem of how to treat the high rate of missing values, as is usual in this type of scale. **Method:** First, responses to the EPV-R in a sample of 1215 male abusers who were reported to the police were used to analyze the patterns of occurrence of missing values, as well as the factor structure. Second, we analyzed the performance of various imputation methods using simulated data that emulates the missing data mechanism found in the empirical database. **Results:** The imputation procedure originally proposed by the authors of the scale provides acceptable results, although the application of a method based on the Item Response Theory could provide greater accuracy and offers some additional advantages. **Conclusions:** Item Response Theory appears to be a useful tool for imputing missing data in this type of questionnaire.

Keywords: Abuse, missing values, imputation, item response theory.

Resumen

Comparación de métodos para el tratamiento de valores perdidos en la EPV-R. **Antecedentes:** el desarrollo de un instrumento eficaz para evaluar el riesgo de violencia contra la pareja representa un tema de gran relevancia social. En el presente estudio se evalúa la escala de Predicción del Riesgo de Violencia Grave Contra la Pareja –Revisada– (EPV-R), una herramienta desarrollada en nuestro contexto, que se enfrenta al problema de cómo tratar la elevada tasa de valores perdidos, que es usual en este tipo de escalas. **Método:** en primer lugar, se estudia en una muestra empírica (N = 1215) el patrón de aparición de los valores perdidos, así como la estructura factorial del EPV-R. En segundo lugar, se analiza el funcionamiento de distintos métodos de imputación en datos simulados en los que se emula el mecanismo de pérdida de datos encontrado para la base de datos empírica. **Resultados:** el procedimiento de imputación originalmente propuesto por los autores de la escala ofrece resultados aceptables, si bien la aplicación de un método basado en la Teoría de la Respuesta al Ítem podría proporcionar una mayor precisión y ofrece algunas ventajas adicionales. **Conclusiones:** la Teoría de la Respuesta al Ítem demuestra ser una herramienta útil para la imputación de respuestas en este tipo de cuestionarios.

Palabras clave: maltrato, valores perdidos, imputación, teoría de la respuesta al ítem.

Missing data is a common problem in research with questionnaires (Van Ginkel, Sijtsma, Van der Ark, & Vermunt, 2010). In this study we aim to compare different imputation methods for treating missing responses on a specific tool, derived from the "Severe Intimate Partner Violence Risk Prediction Scale" (EPV, Echeburúa, Fernández-Montalvo, Corral, & López-Goñi, 2009). This scale has 20 items and is hetero-applied to partner aggressors reported in the Basque country as of 2007. However, a problem of this questionnaire was the presence of missing values, questions that are left unanswered due to lack of information. Echeburúa, Amor, Loinaz, and Corral (2010) proposed a revised scale (EPV-R) to address this problem with indications of how to deal with the missing items. These authors reported an internal consistency of .72 and an interrater reliability for the scale scores of .73.

Received: March 1, 2016 • Accepted: March 3, 2017
Corresponding author: Francisco J. Abad
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: fjose.abad@uam.es

For scoring the EPV-R, Echeburúa et al. (2010) proposed a method to evaluate incomplete questionnaires and to impute the final risk scores. They classified 20 items into three groups according to their discrimination (corrected item-total correlation): low ($r \leq .19$), medium ($.19 < r < .27$) and high ($r \geq .27$). In general terms, their proration method consists of imputing as a function of the response pattern of the scale items completed in each case, which depends on the available information from the victim or the offender. The values were imputed separately for each block of items (low, medium or high discrimination). If the number of missing values was small, the score imputed in the block was set to be proportional to the score obtained in the completed items of that block. If there were a large number of omissions, the score obtained in the completed items of the high-discrimination block was taken as reference to impute the medium- or low-discrimination blocks. Lastly, if the number of missing values was too high, the profile was considered null. After imputing the values, they calculated the total score, weighting the scores in the items of low, medium, and high discrimination, respectively, by one, two, and three.

One of the main motivations of this study is the complex task involved in using the proration method proposed by Echeburúa

et al. (2010): the items must be considered separately according to their discrimination, and the total score of each subgroup of items and the number of missing values must be calculated. This process is not only complex, but also the factor of human error in the imputation may introduce additional errors in the application of the method when the scoring is not automatically automatized.

On the other hand, the proposed proration method is similar to person-mean imputation, whose limitations are well known. Schafer and Graham (2002) suggest that this procedure can lead to biased estimates. A potential problem of this technique is that the imputation is made through different items for each person. Therefore, the homogeneity of the items regarding their means and discrimination is a variable used to measure the efficiency of the method (Enders, 2010). In this sense, as in the EPV-R imputation is made by separating items according to blocks of discrimination, some homogeneity in discrimination is expected for the items within each block. However, the items within each block may be heterogeneous in terms of their means and thus the person's score may still depend on the completed items.

There is no study on the performance of the proposed method. This paper analyzes it in comparison with other more traditional approaches and one approach based on the Item Response Theory (IRT). To perform this comparison, it is necessary to have a realistic model on how responses and missing values on this scale occur, so firstly, we examined the fit of the model of Holman and Glas (2005) for the treatment of non-ignorable omissions.

Missing item scores are usually classified (e.g., Enders, 2010) as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). If missing item scores are MCAR, the presence of missing is unrelated with the measured variables and the observed responses constitute a simple random sample from all scores in the data (e.g., when an applicant accidentally skips an item). If missing item scores are MAR, the presence of missing is related to one or more observed variables (e.g., the information regarding the past behavior can be less available when the aggressor has a foreign origin). Finally, missing responses on an item are MNAR when the probability of missing it relates to the values of the item itself (e.g., when the victim feels afraid or embarrassed and fails to report the most serious threats). In the latter case, the missing data are non-ignorable and specifying a model for the missingness might be needed to achieve a good performance.

Holman and Glas (2005) propose a two-dimensional IRT model for the treatment of non-ignorable missing values. In this model, two indicators are established for each item *j*: Y_j (which represents the subject's response) and d_j (which indicates the presence of a missing value). In this factor model, the d_j variables come from a factor of the propensity to omit (θ_2), and the Y_j variables derive from the level of the trait on which the observed responses depend, for example, the level of risk (θ_1). This model is represented in Figure 1 for a three-item questionnaire.

In the model of Figure 1, the correlation between θ_1 and θ_2 would indicate the degree of ignorability of the missing values: the farther away from zero, the less the ignorability. For example, a positive correlation would imply that the presence of missing values tends to be more frequent in the people at more risk, θ_1. In that case, we would refer to a MNAR missing data mechanism because the presence of a missing value in the variable Y_j depends on the values of Y_j, even when controlling for the rest of variables observed in the analysis. Therefore, if we ignored the framed part of the model (the missing data mechanism), the maximum-likelihood estimates of the model parameters would be biased.

Some of the classical imputation methods are based either on the person-mean or on the item-mean imputation. With respect to person-mean imputation, even though it might be recommended for its simplicity for unidimensional scales (Fernández-Alonso, Suárez-Álvarez, & Muñiz, 2012) presents multiple limitations, as outlined earlier. Item-mean imputation is not recommended either, because it does not consider the results of the person, it reduces the variability of the analyzed variable, and it can affect the relations with other variables (Schafer & Graham, 2002). An intermediate strategy is the two-way imputation, which takes into account the results per item and per person and which has obtained suitable results as an imputation procedure for questionnaire responses (Bernaards & Sijtsma, 2000).

Another possibility is to use the available information in the response pattern of the completed items, such that the responses given by the person and the characteristics of the completed items are considered. This can be done using IRT models. One of the IRT-based methods of imputation is proposed by Hardouin, Conroy, and Sébille (2011). This method calibrates the database with missing values using the corresponding IRT model. It uses the estimated parameters of the items and the person to calculate the probability of a value after applying the IRT model function. This probability can be dichotomized using Bernoulli's function.

The goal of this paper is twofold. First, we will analyze in an empirical sample: (a) the degree to which a one-dimensional IRT model can be applied to the observed data; (b) the internal structure of the missing values; and (c) the possibility of applying the model of Holman and Glas (2005) to establish the degree of ignorability of the omissions. In this way, we will establish a model of a realistic description of the missing data mechanism and will determine whether there is sufficient one-dimensionality to apply an IRT model.

Second, we will compare the following imputation methods in a simulated dataset: (a) Item-mean imputation (b) Person-mean imputation; (c) Two-way imputation; (d) IRT imputation: Imputing the probability of "say yes" according to the IRT model (i.e., the missing response is substituted by $P[Y_j=1|\theta]$); and (e) Imputing with the method proposed originally for the EPV-R. It is expected that the best imputation method will be the one based on IRT, as it takes into account the characteristics of the subjects and the items.

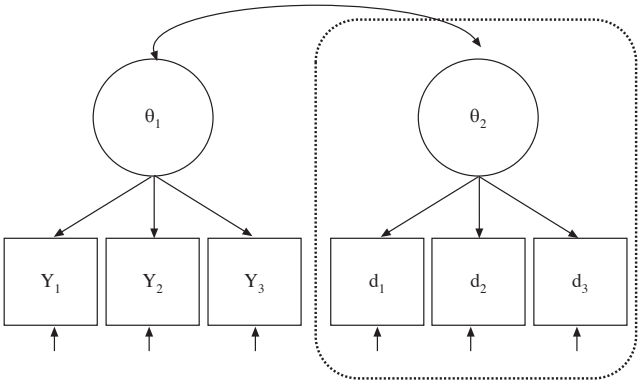

Figure 1. A general bi-dimensional model for the treatment of missing values. θ_1 indicates the measured latent trait; Y_1, Y_2 and Y_3 indicate the observed responses to the items; θ_2 indicates the latent propensity for missing; d_1, d_2 and d_3 indicates the presence of missing responses. The framed part of the model indicates the missing data mechanism

Method

Participants

We examine the internal structure of the observed responses and of the missing values in a sample of 1215 aggressors. The inclusion criteria were: (a) have been denounced for applying violence against a partner or former partner; (b) meet the criteria of the authors of the scale for their scores to be considered valid (responses to at least 12 items and at least 6 responses to the 11 most discriminative items). The latter criterion is consistent with the applied in other similar instruments as the ODARA (Hilton et al., 2004). In our case, we excluded 647 protocols out of 1862.

Instruments

The EPV-R was the scale used. The EPV-R consists of 20 items related to risk of gender violence by men against women.

Procedure

The police interviewed all the participants at the time of the complaint. A supervisor checked the evaluations in order to guarantee response accuracy. The responses on the EPV-R are dichotomous (Yes = 1/No = 0). We additionally create new variables in which we coded the absence (0) or presence (1) of the missing value in each item.

Data analysis

Data generation. The simulated dataset was generated according to the model of Holman and Glas (2005) applied to the empirical dataset. For the items, we used the parameters estimated in the empirical dataset. We simulated the responses of 10000 people from a bivariate standardized normal distribution of trait level (θ_1) and propensity for missing (θ_2) with $r(\theta_1, \theta_2) = .10$. First, we generated the complete data matrix from θ_1 and, subsequently, the missing values were generated from θ_2. For the simulated data, the average rate of missing values per item ranged from .01 to .32 ($M = .11$; $SD = .09$) and the average rate per simulee ranged from 0 to .7 ($M = .11$; $SD = .13$). The 66% of simulees had one or more missing responses.

We applied each of the imputation methods to each incomplete database generated. In the case of the EPV-R imputation method, we applied the method of Echeburúa et al., (2010). In the case of the IRT imputation method, we assumed the logistic two-parameter model estimated in the empirical sample that met the selection criteria ($N = 1215$) and obtained weighted maximum likelihood trait θ_1 estimates (Warm, 1989). The missing values in each item were replaced by the probability predicted according to the IRT model in that item. In the case of the item-mean imputation and person-mean imputation methods, the missing values were replaced by the item mean, or the person mean, respectively. In the case of item-mean imputation, we imputed by the mean of the item obtained in the empirical database with complete response patterns ($N = 450$). In the two-way imputation, the score was obtained as: person mean + item mean – overall mean, where overall mean is the global mean across all the responses.

In all the applied procedures, imputation was made at the item level, and the final score obtained is the weighted sum of the items.

Data analysis in the empirical dataset. Firstly, we separately analyzed the assumption of unidimensionality of the observed responses and the missing values, following the parallel analysis procedure and the Hull method, both of them implemented in the FACTOR program (Lorenzo-Seva & Ferrando, 2013) and applied to the matrix of tetrachoric correlations. The ratio of the first two eigenvalues was also obtained (Gorsuch, 2003).

Secondly, we applied the two-parameter logistic IRT model for observed responses (Y) and missing presence (d) variables. To estimate item parameters, we used a Bayesian marginal maximum-likelihood estimation procedure. The means of the prior distributions for the item parameters were fixed by an iterative procedure based on empirical data. Item fit was analyzed by the Orlando and Thissen (2000) χ^2 statistics. Lastly, we applied the two-dimensional IRT model of Holman and Glas (2005) through the robust maximum-likelihood estimation procedure (MLR), which allows the inclusion of missing values.

The application of the IRT models was done through the mirt package (Chalmers, 2012). The application of the model of Holman and Glas (2005) was done with Mplus (Muthen & Muthen, 2012).

Data analysis in the simulated dataset. The score obtained with the complete response pattern, without missing, was considered the "true" score. For each method of imputation, we considered as precision measurements: (a) the correlation between the score obtained with the complete response pattern (X_{emp}) and the score imputed by the method (X_{imp}); (b) the mean absolute error or difference between X_{emp} and X_{imp} (*MAE*); and (c) the mean bias. Regarding the bias, the results for each trait level (θ_1) were also considered, distinguishing 8 levels of θ_1 (taking as cut-off points the values included between -1.5 and 1.5 in intervals of 0.5 points).

Results

Application of the Holman and Glas model to the empirical dataset

The dataset with the observed responses and the dataset with the missing presence variables were analyzed separately. For both the cases, the criteria suggested a structure of between one (Hull method) and two factors (parallel analysis), the ratio between the first and the second factor was greater than 2, and a dominant general factor explained an important part of the common variance (27.5% and 29.5%, respectively). Given these results, we considered that a one-dimensional IRT model was applicable in each case. The presence of a general factor for missing presence variables implies that, when there are missing values in one item, there are usually also missing values in other items. When applying the IRT one-dimensional model, we found that practically all the items adequately fit the model ($p>.05$), except for two items in each case.

The standardized loadings observed in the bi-dimensional model of Holman and Glas (2005) are similar to those obtained when each model is applied to each type of responses separately. The correlation between the two columns of loadings was positive, but not statistically significant ($r = .372$, $p = .106$; see scatter plot in Figure 2). In fact, when deleting the first item, the correlation decreased to .122 ($p = .619$). There was a high correlation between the item standardized loadings on the omission propensity factor and the percentage of missing values in the item ($r = .816$, $p<.001$; see scatter plot in Figure 3). This implies that, when people tend

to skip one item, they tend to skip others, but especially in the case of items with more omissions (e.g., items 12, 13 and 15, that refer to the past behavior of the aggressor). At the person-level, the correlation between θ_1 and θ_2 was statistically significant but small (r = .101, p = .022).

Accuracy of response imputation methods

The precision outcomes for each imputation method showed that the highest levels of precision were usually obtained for the IRT imputation procedure, followed by the two-way imputation procedure, the person-mean imputation method, the score proposed by the authors of the scale (EPV-R imputation), and finally, the item-mean imputation method, although the differences were small (see Table 1). The main effect of the method on the *MAE* was statistically significant, $F(4, 6245) = 124.5$, $p<.001$, $\eta^2 = .074$. Comparisons among methods were significant ($p<.001$), except between the EPV-R imputation method and person-mean imputation method ($p = .188$). The largest standardized difference was small and was found between the IRT imputation method and the item-mean imputation methods, Cohen's $d = -0.187$.

Figure 4 shows the bias for each level of θ_1. The mean bias of the item-mean imputation has the largest absolute value and it changes from positive to negative as the trait level increases. In relation to the rest of procedures, we observed that, at low levels of trait, the procedures provide very similar results, whereas at high levels, the bias of the EPV-R imputation procedure tends to increase. In general, the differences in bias between the IRT imputation and the EPV-R imputation become more pronounced as the subject's trait level increases. A similar result can be concluded from the comparison of the person-mean imputation and the EPV-R imputation. In general, it seems that the effect of

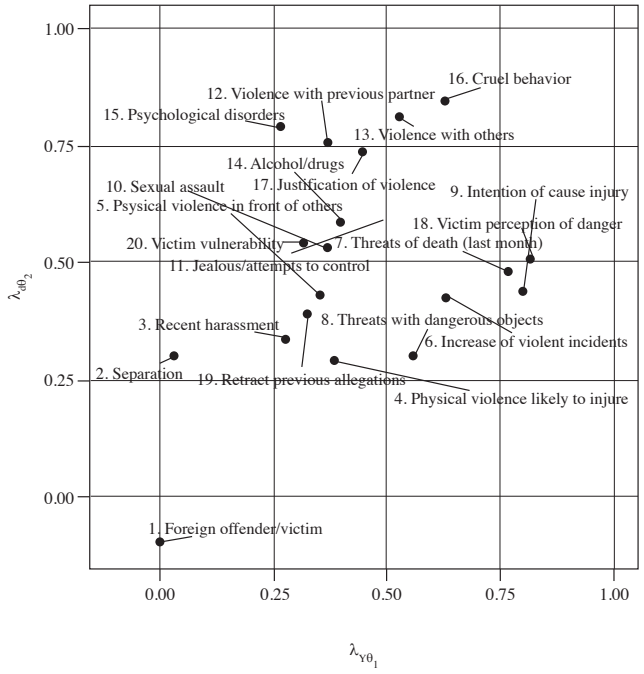

Figure 2. Scatter plot of standardized factor loadings of the *Y* variables (observed responses) on the latent Risk factor (θ_1) and standardized factor loadings of the *d* variables (missing presence variables) on the latent propensity for missing factor (θ_2)

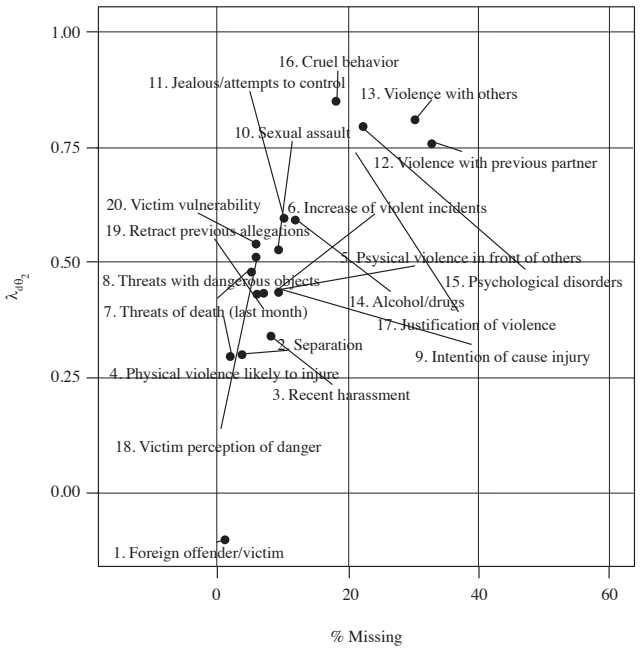

Figure 3. Scatter plot of Percentage of omissions (% Missing) and standardized factor loadings of the *d* variables (missing presence variables) on the latent propensity for missing factor (θ_2)

Table 1
Bias, MAE, and correlation between the Score with complete data (X_{emp}) and the imputed score (X_{imp}) according to each imputation procedure

Response imputation method	Bias		MAE		r
	M	SD	M	SD	
IRT	0.06	2.04	1.50	1.39	.977
Item-mean	-0.27	2.39	1.78	1.61	.973
Person-mean	0.09	2.20	1.65	1.46	.974
Two-Way	0.02	2.13	1.57	1.44	.975
EPV-R	0.20	2.50	1.67	1.87	.970

Note: n = 6,249 out of 10,000; simulees with a complete response pattern or with an invalid protocol according to the EPV-R criteria were excluded

imputing according to the mean separating by blocks has a negative effect versus an imputation based on the general questionnaire. This negative effect tends to occur when the mean of positive responses in the most discriminative block is significantly higher than the mean of positive responses in the other blocks. As the EPV imputation procedure grants an excessively higher weight to the discriminative block, a positive bias occurs.

Table 2 allows comparing the classification errors (in relation to the complete response pattern) for all the tested imputation methods. IRT, two-way and person-mean imputation response methods performed in a similar way. Cohen's κ coefficients of agreement with the true classification were .839, .832 and .82, respectively. For the EPV-R imputation, Cohen's κ was .829. The differences among these methods were small. When comparing them, it can be seen that men were more easily classified at the high risk level with EPV-R. This is due to the score positive bias, and, consequently, EPV-R imputation overestimates the risk

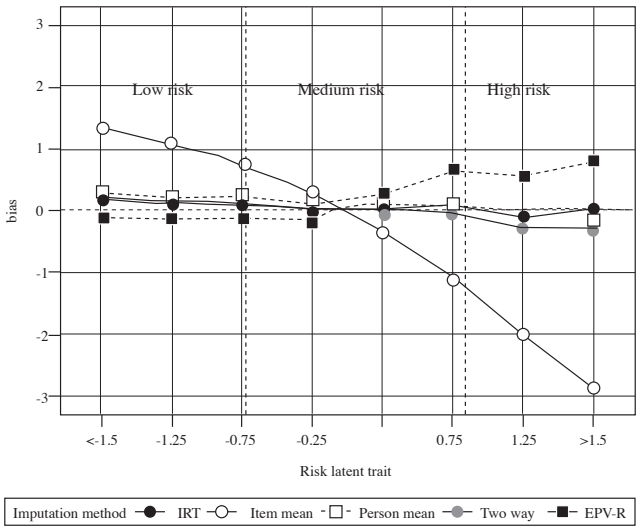

Figure 4. Bias as a function of the Risk latent trait (θ_1) for each imputation procedure

at the medium level, being less accurate (e.g., IRT, 89.6%, vs. EPV-R, 84.4%). Finally, the item-mean imputation method had the worst classification rate, with a κ coefficient of agreement of .798. Item-mean imputation method increased the classification in the medium risk level, but at the expense of decreasing the classification accuracy for the remaining risk levels.

Discussion

Assuming the model of Holman and Glas (2005), a positive significant but low correlation emerges between the latent level of "aggressor violence risk" and the latent level of "propensity to present missing values in the responses". Therefore, the missing values may be non-ignorable. This could be due to the fact that a high score on the scale could have negative consequences for the subject and, therefore, the potential aggressor will tend to omit more questions, especially those that more clearly identify him as having an offender profile. Sometimes, the missing value may have to do with the fact that the informant is exclusively the victim and that the interviewer has no information about that datum, especially when it refers to historical variables. Therefore, when attempting to impute missing values, their non-ignorability should be taken into account using a suitable model, as suggested by Huisman and Molenaar (2001). We propose to use the model of Holman and Glas for the study of missing values. However, taking into account the small size of the correlation, we consider relevant investigate the use of simpler models that treat missing as ignorable.

In general, missing values tend to appear together (the presence of omissions in these items tends to covary), according to their loading on the general factor of propensity for missing. The items with the highest loading on this factor are those referring to the history of behaviors (Items 12 and 13), the existence of a psychiatric history (Item 15), the presence of cruel behavior and contempt (Item 16), and the justification of violent behavior (Item 17). These items tend to be the ones with the most missing values (over 18%). However, there is no significant relationship between the proportion of missing values in these items and the severity of the violent behavior reflected in the item or its discriminative ability for risk prediction.

The results of the simulation study indicate that there are small differences between the proposed imputation methods. We found evidence supporting that the simplest imputation, item-mean imputation, produces greater MAE and bias, so we do not recommend its use. This recommendation is consistent with the conclusions provided by other authors for the treatment of missing values (Schafer & Graham, 2002). Regarding the rest of the procedures, the more information used by the procedure, the better it performs. The order of the procedures, from best to worst performance, would be: IRT imputation, two-way imputation and person-mean imputation. The original proposal (EPV-R imputation) produces the worst results, especially at the levels of medium-high risk when evaluating bias as a function of trait. This has some impact on the classification of medium trait levels, which is important, because the scale is administered to people reported for abuse, and therefore it can be assumed that it targets individuals with medium-high trait levels, where the IRT or two-way imputation method produces lower bias in raw scores. On the other hand, the classification rates are so similar for the IRT, two-way and person-mean methods, so that the latter methods might be recommended by their simplicity. All the procedures can be programmed easily with a software.

In the current study we have focused on single imputation methods, applied for scoring a specific instrument. We acknowledge that multiple imputation methods or maximum likelihood could be preferable for dealing with missing data (Enders, 2010). Here we appraised the implementation of multiple imputation, but we needed an imputation that would assign a unique score to each subject. However, single imputation does not take uncertainty about the missing data into account, so that standard errors of statistics will be biased downward. In this regard, we would

Table 2
Contingency tables between the classification obtained with complete response pattern (True risk) and those obtained through the imputation procedures (Percentages by rows)

True risk	Estimated Risk					
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
	IRT			**Item-mean**		
Low[a]	89.7%	10.3%	0.0%	80.6%	19.4%	0.0%
Medium[b]	5.6%	89.6%	4.9%	3.5%	94.7%	1.8%
High[c]	0.0%	9.5%	90.5%	0.0%	18.8%	81.2%
	Person-mean			**Two-way**		
Low[a]	88.1%	11.9%	0.0%	89.5%	10.5%	0.0%
Medium[b]	6.2%	88.3%	5.6%	5.8%	89.1%	5.2%
High[c]	0.0%	10.1%	89.9%	0.0%	10.1%	89.9%
	EPV-R					
Low[a]	92.8%	7.2%	0.0%			
Medium[b]	7.8%	84.4%	7.8%			
High[c]	0.0%	5.7%	94.3%			

Note: n = 6249 out of 10000; simulees with a complete response pattern or with an invalid protocol according to the EPV-R criteria were excluded
[a] n = 1692; [b] n = 3045; [c] n = 1512

like to highlight that maximum-likelihood IRT estimates can be directly used for classification: the IRT framework does not require imputation. Indeed, this alternative IRT approach has an advantage over tested procedures because it provides an indicator of the standard error of measurement which, as usual, will depend on the trait level but also on the number of missing values. This is important, because the confidence interval for each score is provided. Finally, it should be noted that IRT could also be useful for generating multiple imputations if statistical analysis are required and we need taking into account the uncertainty in the imputation process.

Acknowledgements

The research has been funded by the Ministry of Economy and Competitivity of Spain, project PSI2013-44300-P.

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321-364.

Chalmers, R. (2012). Mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

Echeburúa, E., Amor, P. J., Loinaz, I., & Corral, P. (2010). Escala de Predicción del Riesgo de Violencia Grave contra la pareja –Revisada– (EPV-R) [Severe Intimate Partner Violence Risk Prediction Scale-Revised]. *Psicothema*, *22*(4), 1054-1060.

Echeburúa, E., Fernández-Montalvo, J., Corral, P., & López-Goñi, J. J. (2009). Assessing risk markers in intimate partner femicide and severe violence: A new assessment instrument. *Journal of Interpersonal Violence*, *24*(6), 925-939.

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

Fernández-Alonso, R., Suárez-Álvarez, 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.

Fernández-Montalvo, J., Echeburúa, E., & Amor, P. J. (2005). Aggressors against women in prison and in community: An exploratory study of a differential profile. *International Journal of Offender Therapy and Comparative Criminology, 49*(2), 158-167.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology: Research methods in psychology, vol. 2* (pp. 143-164). Hoboken, NJ: John Wiley & Sons.

Hardouin, J., Conroy, R., & Sébille, V. (2011). Imputation by the mean score should be avoided when validating a patient reported outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Medical Research Methodology*, *11,* 105.

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario domestic assault risk assessment. *Psychological Assessment*, *16*(3), 267-275.

Holman, R., & Glas, C. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical & Statistical Psychology*, *58*(1), 1-17.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response theory. In A. Boomsma, M. Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 221-244). New York: Springer.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). Factor 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Applied Psychological Measurement*, *37*(6), 497-498.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus Version 7: User's guide*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

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

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.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.