

LI Analysis Training Series

Imputation of Index Scores with Missing Data

(Last Revised: 4/23/07)

Rodney Funk, Melissa Ives and Michael Dennis

Chestnut Health Systems

Bloomington IL 61701

309-827-6026

www.chestnut.org

Acknowledgement:

This document was developed under contract #270-2003-00006 from the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). Any opinions about this data are those of the authors and do not represent official positions of the government or individual grantees.

Purpose: To provide a preferred method for calculating values for summative index scores with missing data. Examples for four different indices will be shown with equations derived from regression analysis. The methods described herein are based on data cleaning done with the Global Appraisal of Individual Needs (GAIN; Dennis, Titus, White, Unsicker, & Hodgkins, 2003).

Background: The GAIN includes over 100 computed scales and indices. Scales follow the classical measurement model where internally consistent items are combined to estimate an underlying trait. Since these scales have a high inter-item correlation and Cronbach's alpha $>.7$, as long as there are 3 or more valid answers to a scale, the scale is computed from the average of the valid items. To see more of an explanation on how these scales are calculated, see the *Data Cleaning and Replacement of Missing Values* (McDermeit [now Ives], Funk, Dennis, 1999; available at www.chestnut.org/li/downloads/...

Index scores, however, do not require items to be internally consistent with one another. These indices have low inter-item correlation and alpha are often less than $.7$. They are computed by summing over all items, so valid answers to all the items in the index are required. If one item of an index is missing, then the index score will not be created and set to system missing. This can lead to a lot of missing data, especially if one of the items in the index is not part of the core GAIN (required item). For any analysis performed with an index and using listwise deletion, this could cause a reduction in statistical power (Dennis, Lennox, & Foss, 1997) and could lead to biased estimates (Little & Rubin, 1987). Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. It is therefore recommended that data be replaced in one or more of the advanced methods based on hot-deck imputation (McDermeit [now Ives], Funk, Dennis, 1999), multiple imputation (modeling uncertainty due to missing data, while using the existing data (Rubin, 1987)) or a regression model (predicting the missing value based on the other available data). The following examples used the regression model.

Procedure: Since the items in indices are not highly inter-correlated, we only replace an item when there is only one missing. The GAIN also has certain items that are required when used, but not all the items are required. When there is one item missing due to it systematically not being asked by site choice (which by definition is “completely at random” relative to the individual), we will impute the one item not asked and no more than one other non-systematically missing item. These items are imputed using regressions based on the other items used to create the index. We used data from over 9,000 adolescents interviewed with the GAIN-I version 5, as part of 89 CSAT adolescent treatment grants to develop the regression equations.

To establish an imputed value, we used several steps. First, we established the minimum number of items required to impute an equivalent score as no fewer than 11 and preferably 12 of 13 items. Then we created regression equations for each index item using the remaining items. The appropriate equation (selected using syntax) is then used to impute the index value for records where only one item was missing.

Example 1 Recovery Environmental Risk Index (RERI).

The example below used the Recovery Environmental Risk Index (RERI). It is based on the sum of 13 items that include days or recency (divided by range) of AA attendance (reversed), homelessness, living with alcohol or drug use in the home, violent arguments, and physical, emotional or sexual abuse. This sum is then divided by 13 so that the index runs from 0 to 1 with higher scores indicate less involvement in support groups (e.g., AA, CA, NA) and more environmental risk from alcohol/drug use in the home, fighting and/or victimization or proportionally more recovery environmental risk for relapse. This index has been found to be related to and predictive of social environmental risk, substance frequency and substance use (Godley, Kahn, Dennis, Godley, and Funk, 2005). This index makes a good example since a) items come from several sections of the GAIN instrument, b) it usually has a high (more than 5%) amount of missing data and c) it has one item that is not required and can be systematically not asked. When there are valid responses to all 13 items in the RERI, this syntax computes the score:

```
compute rs6ap=1-s6a/90.
compute e1bp=e1b/6.
compute e1dp=e1d/90.
compute e2cp=e2c/90.
compute e2dp=e2d/90.
compute e2ep=e2e/90.
compute e3p=e3/90.
compute e8_p=e8/6.
compute e8pp=e8p/90.
compute e9tp=e9t/6.
compute e9up=e9u/90.
compute re14ap=1-e14a/90.
compute e14bp=e14b/90.
compute reril3p=sum.13(rs6ap to e14bp)/13.
```

If any one of the above 13 items is missing, then the index score (rer13p) is left as system missing. Item E2c above is not one of the core items in the GAIN, so this item may not be asked if a program or site chooses to omit it.

Imputation:

We had valid answers to all 13 items for only 2565 cases; less than a third of the records. This was mainly due to missing the optional item E2c. In our first attempt to impute index scores, we tried to see how many items were required to get an equivalent score to use the average of the valid responses to impute the scale score. This is the same type of process we use in computing classical scale scores. We selected items randomly from those cases with all 13 items and imputed the index score starting with 3 valid answers and working our way up to only one missing value. This process was repeated for three other index scores: the Social Risk Index, the Personal Sources of Stress Index and the Other Sources of Stress Index. We then compared the means of the imputed values with the mean when using all 13 items. We found that after 1 or 2 missing items, the effect sizes for the differences in the means were approaching small in size, $d=0.2$. This suggested that imputation should ideally only occur with only 1, or at the most 2 missing items.

Our next step was to try to impute the missing item using the remaining items used to create that index. Due to the finding above, we limited the imputation to when there was only one missing item. The first step was to run regressions to predict each item used in RERI using the other items in the index, for example:

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT rs6ap  
/METHOD=ENTER e1bp e1dp e2cp e2dp e2ep e3p e8_p e8pp e9tp e9up re14ap  
e14bp
```

There would be one of the above used for each item in the index making the appropriate adjustments to the /DEPENDENT and /METHOD subcommands until each item had been run as the /DEPENDENT variable. The coefficients from the output would then be used to create equations to impute the items when they are missing. These would be found in the coefficients table under the unstandardized coefficients, column 'B'.

Coefficients^a

Model	el	Unstandardized Coefficients		Standardized Coefficients		Collinearity Statistics		
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.965	.006		166.742	.000		
	e1bp Proportion of Recency of homelessness/Range (6)	-.025	.016	-.035	-1.557	.120	.776	1.288
	e1dp Proportion of days homeless/Past 90	-.006	.029	-.005	-.212	.832	.778	1.285
	e2cp Proportion of days in shelter -emerg hsg./Past 90	-.216	.048	-.091	-4.534	.000	.952	1.051
	e2dp Proportion of days anyone used alcohol at home/Past 90	.010	.005	.041	1.944	.052	.876	1.142
	e2ep Proportion of days anyone used drugs at home/Past 90	.007	.008	.019	.910	.363	.844	1.184
	e3p Proportion of days in trouble w/family/Past 90	.015	.007	.046	2.149	.032	.846	1.181
	e8_p Proportion of Recency last argue-threat-fight/Past 90	.000	.006	-.001	-.051	.959	.768	1.303
	e8pp Proportion of days of arguing/Past 90	-.016	.010	-.039	-1.673	.094	.710	1.408
	e9tp Proportion of Recency: last attacked-abused/Range (6)	-.020	.009	-.053	-2.282	.023	.701	1.427
	e9up Proportion of days attacked-abused/Past 90	.048	.019	.057	2.506	.012	.738	1.355
	re14ap REVERSED Proportion of days in activity-no one using AOD/Past 90	.017	.006	.058	2.936	.003	.978	1.022
	e14bp Proportion of days in activity-anyone using AOD/Past 90	.013	.009	.029	1.424	.154	.955	1.047

^a. Dependent Variable: rs6ap REVERSED-Proportion of days attend self-help groups/Past 90

The next step was doing the imputation for the missing items. We add a prefix 'm' for variables where missing data has been replaced or imputed. First, we create these new variables from the existing variables:

```
compute mrs6ap=rs6ap.
compute me1bp=e1bp.
compute me1dp=e1dp.
compute me2cp=e2cp.
compute me2dp=e2dp.
compute me2ep=e2ep.
compute me3p=e3p.
compute me8_p=e8_p.
compute me8pp=e8pp.
compute me9tp=e9tp.
compute me9up=e9up.
compute mre14ap=re14ap.
compute me14bp=e14bp.
exe.
```

This is important so that when the index score is recalculated, it will be the same for those who were not missing any RERI item. Then, where an item is missing, it is imputed using the regression equation:

For example, a record missing only S6a could impute mrs6ap¹ for use in calculating the imputed index mRERI13p.

```
do if rivld=12 and missing(mrs6ap).
  compute mrs6ap=(0.965488864556224+e1bp*-0.0252060008088797+e1dp*-
  0.00615777177201543+e2cp*-0.215601281265844+
  e2dp*0.010202666613149+e2ep*0.00735607742064711+
  e3p*0.0149561295556361+e8_p*-0.000311539996538543+e8pp*-
  0.0162152343880479+e9tp*-0.0198088180702741+
  e9up*0.0475481499103086+re14ap*0.0170497595771667+
  e14bp*0.0131087479000485).
end if.
```

The above compute statement was derived from pasting the coefficients table above into Excel. The concatenate function in Excel is used along with the coefficients to create the compute statement. The compute statement is then pasted back into the SPSS syntax (see Discriminant analysis memo (Ives, Funk & Dennis, 2007) for more details). The 'rivld' is the count of valid items for RERI², so the imputation will only be done when there is only one missing item from the index. So 13 equations would be created in all, one for each of the items in RERI.

As mentioned before, there was also one item, E2c, which was systematically missing since it is not being a required item. So we also imputed the missing item if E2c and only one other item were missing with the following syntax:

```
do if rivld=11 and missing(mrs6ap) and missing(me2cp).
  compute mrs6ap=(0.94997542517717
  +e1bp*-0.0469508211058038
  +e1dp*0.00764545998280615
  +e2dp*0.0103499891557881
  +e2ep*0.00361560024562884
  +e3p*0.0201777535792262
  +e8_p*-0.00136872860603936
  +e8pp*-0.00579173607561245
  +e9tp*-0.0223818483407785
  +e9up*0.0257689358461028
  +re14ap*0.0309993451717349
  +e14bp*0.0131366577622224).
end if.
```

¹ This name (mrs6ap) incorporates 3 of our standard naming conventions: 'm' indicates that this is the missing replaced version of the variable; 'r' indicates that the original variable was reversed to match the direction of the index (higher risk=higher value); s6a is the original variable name; 'p' indicates that the item has been divided by it's range to create a proportional (0-1) variable.

² Calculated using the 'nvalid' SPSS function (or can use syntax)

Once the imputation of missing items was completed, we then created the index in the original manner using the imputed variables. This works because we started by setting each imputed item to the value of the existing item and only imputed any remaining missing values where there was only one missing item in the scale or where E2c and one other item were missing. If these criteria were not met (i.e. missing >2 items), the index remained missing.

```
compute mreri13p=sum.13(mrs6ap to me14bp)/13.
var labels mreri13p
      'recovery environmental risk index, using regression for missing an item'.
desc mreri13p.
```

Example 2 Personal Sources of Stress Index (PSSI)

The above example used only items that were continuous measures. For dichotomous yes/no items (1/0), a logistic regression would need to be used to get the coefficients for imputation. For example, the Personal Sources of Stress Index (PSSI) is the sum of 6 yes/no items (E10_1 to E10_99). The regression would be:

```
LOGISTIC REGRESSION E10_1
/METHOD = ENTER E10_2 E10_3 E10_4 E10_5 E10_99
/PRINT = GOODFIT CI(95)
/CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

The compute statements for the imputation are calculated using the logistic regression coefficients to calculate the probability of answering yes to that item. The compute statement is rounded so values <0.5 become 0 and >0.5 become 1. For E10_1, the equation would be:

```
do if pssivld=5 and missing(me10_1).
  compute me10_1=rnd(1/(1+(exp(-(-3.10958505822322
    + E10_2*0.828496639599022
    + E10_3*0.336460465395659
    + E10_4*0.5236773156897
    +E10_5*0.0615586230260587
    + E10_99*0.397246323135209)))))).
end if.
```

Once again, the index score is calculated then from the imputed items³:

```
compute mpssi=sum.6(me10_1 to me10_99).
var labels
  mpssi 'Personal Sources of Stress Index, using regression for missing an
  item'
  pssi 'Personal Sources of Stress Index'.
freq pssi mpssi.
```

³ which were set to the existing item values prior to any imputation.

Comments: Using the above imputation at the item level increased the number of valid RERI scores to up over 8000 (<5% missing). We checked correlations of the original RERI score and the imputed version (MRERI) with the 4 core GAIN factor measures and some of the most used change scores used in GAIN analysis. The correlations were very similar for both methods of computing RERI with no significant differences by method of calculation.

Correlations

		rer13p Recovery Environment Risk Index	mreri13p2 recovery environmental risk index, using regression for missing an item
Pearson Correlation	giss General Individual Severity Scale	.510	.501
	spsy Substance Problem Scale (Past Year)	.293	.291
	imds Internal Mental Distress Scale--4 factors	.405	.410
	bcs Behavior Complexity Scale	.426	.424
	cvs Crime & Violence Scale--4 factors	.463	.452
	sri7 Social Environmental Risk Index--New (V5)-7 items	.307	.339
	SFS8p Substance Frequency Scale 8-item version	.363	.396
	spsm Substance Problem Scale (Past Month)	.348	.364
Sig. (2-tailed)	giss General Individual Severity Scale	.000	.000
	spsy Substance Problem Scale (Past Year)	.000	.000
	imds Internal Mental Distress Scale--4 factors	.000	.000
	bcs Behavior Complexity Scale	.000	.000
	cvs Crime & Violence Scale--4 factors	.000	.000
	sri7 Social Environmental Risk Index--New (V5)-7 items	.000	.000
	SFS8p Substance Frequency Scale 8-item version	.000	.000
	spsm Substance Problem Scale (Past Month)	.000	.000
N	giss General Individual Severity Scale	2452	8174
	spsy Substance Problem Scale (Past Year)	2564	8609
	imds Internal Mental Distress Scale--4 factors	2565	8615
	bcs Behavior Complexity Scale	2562	8605

cvs Crime & Violence Scale--4 factors	2565	8617
sri7 Social Environmental Risk Index--New (V5)-7 items	2452	8224
SFS8p Substance Frequency Scale 8-item version	2565	8616
spsm Substance Problem Scale (Past Month)	2564	8609

Describing These Procedures. These procedures would normally be described in a report or paper in the methods section as a way missing data was handled.

Missing data was imputed at the item level using a regression approach (Rubin, 1987, Schaefer and Graham, 2002).

References

- Dennis, M.L., Lennox, R.I., & Foss, M. (1997). Practical power analysis for substance abuse health services research. In K.J Bryant, M Windle, and S.G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research, (pp. 367-405). Washington, DC: American Psychological Association.
- Dennis, M. L., Titus, J. C., White, M., Unsicker, J., & Hodgkins, D. (2003). *Global Appraisal of Individual Needs (GAIN): Administration guide for the GAIN and related measures*. Bloomington, IL: Chestnut Health Systems.
- Figueredo, A.J., McKnight, P.E., McKnight, K.M., & Sidani, S., (2000). Multivariate modeling of missing data within and across assessment waves. *Addiction*, 95 (Supplement 3), S361-S380.
- McDermeit [now Ives], M., Funk, R.R., & Dennis, M.L. 6/24/1999. LI Analysis Training Series: Data Cleaning and Replacement of Missing Values
http://www.chestnut.org/LI/downloads/training_memos/missing_data_1.pdf
- Godley, M. D., Kahn, J. H., Dennis, M. L., Godley, S. H., & Funk, R. R. (2005). The stability and impact of environmental factors on substance use and problems after adolescent outpatient treatment for cannabis use or dependence. *Psychology of Addictive Behaviors*, 19(1), 62-70.
- Little, R., & Rubin, D. B. (1989). The analysis of social science data with missing values. Sociological Methods and Research, 18(2), 292-326.

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

Rubin, D.B. (1996). Multiple imputation after 18 years. Journal of the American Statistical Association, 91, 473-489.

Statistical Program for the Social Sciences. (1997). SPSS Base 7.5 syntax reference guide. Chicago: Author (www.spss.com).