SINRELEF-LD

Score Information, reliability, and relative efficiency in measures that contain locally-dependent items

USER’S GUIDE

Prepared by:

David Navarro-González

Pere J. Ferrando

Fabia Morales-Vives

Please reference this document as:

Navarro-González, D., Ferrando, P.J., & Morales-Vives, F. (2024). *Sinrelef-LD User’s Guide. Technical Report*. Universitat Rovira i Virgili. Tarragona.

Available at: http://psico.fcep.urv.cat/utilitats/SINRELEF-LD/index.html

Contents

1. Theoretical bases and features

Ideally, the items in a scale intended to measure a single dimension of content (e.g. a trait) should be related only because of the influence of the common dimension they measure. In factor-analytic (FA) terms, this condition implies that, once the influence of the common factor has been partialled-out, the item responses cease to be related. When the items that make up a scale meet this condition, they are said to be locally independent. On the contrary, if some of the items continue to be related, they are said to be locally dependent.

In our experience, strict compliance with the local independence condition is more the exception than the rule in personality and attitude measurement. Now, in principle, local dependencies among certain items are generally due to two main reasons (e.g. DeMars, 2020). First is multidimensionality: apart from measuring the "dominant" dimension, some of the items might also measure additional dimensions of content. The second reason is related specificities due to non-content or non-substantive causes, of which, the most commonly considered are: (a) repeated presentation of the same items, (b) content or wording similarities, (c) similarities in the evoked situation, and (d) context effects (Bandalos, 2021, DeMars, 2020, Ferrando & Morales-Vives, 2023). Please note that if item specificities are correlated due to any of these four causes, the sign of the correlation is expected to be positive, because the respondent will tend to answer the items in the same way beyond the influence of the common factor that underlies them. Local item dependencies due to correlated-specificities are the focus of the present program. Furthermore, the program is based on a FA background, in which the units of the analysis are the specific relations between pairs of items. A locally-dependent pair of items that shares specificities is known as a "doublet" in FA terminology. Finally, at this point, a warning note may be useful: In this guide we shall use the terms "local dependences", "correlated residuals", "correlated specificities", or "doublets" indistinctly.

The presence of correlated item specificities that are not taken into account may have negative consequences both when developing a scale or when measuring individuals by using an already existing one. So as to discuss these consequences, we shall use the FA background, and consider the process of scale development and usage as two-stage: Calibration and Scoring (McDonald, 1982). In the calibration stage, the structural items parameters (locations, discriminations/loadings, and error variances) are estimated and goodness of model-data fit is assessed. In the scoring stage, provided that a well-fitting, strong, and replicable solution has been obtained, the item parameter estimates are taken as fixed and known and used as a basis for obtaining individual scores.

At the calibration level, if local dependences exist and are left unmodeled, two main consequences are expected: unacceptable goodness of model-data fit results and biased parameter estimates (see Ferrando et al. 2023 for a thorough discussion). At present, however, and, as discussed below, procedures for fitting a FA solution which includes the residual correlations as additional parameters, are already available. When appropriately used, these "extended" solutions are expected to provide correct calibration results (including fit results) as well as additional information that serves as a basis for the procedures implemented in this package. However, they are less parsimonious and more potentially unstable and prone to capitalization on chance than the "simpler" solutions that assume full local independence.

Turning now to the scoring stage. The package we present is intended to be used with raw (sum) scores based on an extended unidimensional FA solution of the type discussed above. The choice of raw scores is due to several reasons, mainly: (a) They are possibly the most widely used scoring procedure in applications because are easy to compute and interpret, (b) their results are more easily related to those from previous studies, (c) they tend to provide stable results under cross-validation, and (d) they remain unbiased if correlated specificities exist (see e.g. Ferrando et al. 2024 (submitted) or Raykov & Marcoulides, 2011). Even with these advantages (particularly the third one), raw scores based on a "correct" calibration solution might lead to misleading results if the correlated specificities are not taken into account also at the scoring stage. As a summary, if the scores are (wrongly) assumed to be based on a set of fully locally independent items, then, the amounts of accuracy and information they are expected to provide are overestimated (sometimes grossly so), which might have relevant consequences if the scores are used for individual assessment or for individual comparison purposes. Put it in another way: the presence of correlated specificities makes the scores less reliable and informative than they would be if the items were locally independent (Yen, 1993). Unfortunately, in our experience, the reported reliability and information estimates in applications are almost always "optimistically" obtained by assuming full local independence among items. Hopefully, the use of the package we propose can help improve this situation.

1-1. Features implemented in SINRELEF-LD

In this section we shall only provide a conceptual summary of the indices and procedures implemented in the program. Technical details are provided in Ferrando, Navarro-Gonzalez, & Morales-Vives, (2024) to which the interested reader is referred.

All the procedures available in SINRELEF-LD are based on two complementary measures of score accuracy: Information and Reliability. As for the first measure, if the raw scores are considered as estimators of the common factor the test intends to measure, then, the amount of information a raw score provides is inversely proportional to the squared standard error of measurement, and so, to the width of the confidence interval for estimating the 'true' trait level from this score. The amount of information can also be interpreted as a signal/noise ratio: it goes from 0 to infinity, and indicates how many times the common variance of the 'true' trait levels is larger than the error variance in the population of interest.

The coefficient of reliability considered in the program is defined as the (predicted) squared correlation between the raw scores and the 'true' trait levels. From this definition, it follows that this coefficient is the omega reliability coefficient when correlated residuals exist (Bollen, 1989).

The 'key' feature of the proposal is that it computes: (a) the correct information and reliability estimates that take into account the local dependences, and also, (b) the corresponding estimates that would have been obtained if the items analyzed had been all locally independent. The comparison between both sets of estimates gives an idea of the loss of information and reliability that is due to the presence of correlated residuals. To enhance interpretability, we have also proposed a relative efficiency measure, which is the ratio between (a) the correct amount of information estimated by considering the local dependences and (b) the amount of expected information that would have been obtained if the items had been all locally independent. So, a relative efficiency of 0.70 means that the presence of correlated residuals makes the information provided by the raw scores only 70% of the information that the scores would have provided if the items had been all locally independent.

So as to appraise why the reliability and information indices complement each other, and also, how all the results so far summarized are interpreted, we shall consider a numerical example. Suppose that the correct reliability (omega) estimate for the scale scores that correctly accounts for the presence of doublets is ω_{LD}=0.88. Now, suppose that the expected reliability estimate if all the items were locally independent is ω_{LI}=0.92. Clearly, if the items were assumed to be locally independent, the reliability would be overestimated. However, the loss of accuracy due to the local dependences does not seem too great either.

If the differences above are quantified in information terms, however, the picture changes considerably. The information estimates corresponding to the reliability estimates above are now: I_{LD}=7.33 and I_{LI}=11.50, which seems a substantial difference!. The relative efficiency is then: RE_{LD}=7.33/11.50=0.64. So, the information provided by the scores in this item set that contains local dependencies is only 64% of the information that could be attained if these items were all locally independent.

The reason why the information provides a more realistic picture of the loss of accuracy is because this measure is not bounded above. In contrast the reliability is bounded above by 1. So, as we approach one, the reliability estimates are compressed and the differences among them tend to be minimized.

So far, we have discussed a series of indicators intended for the raw scores based on the entire set of items. However, as described below, SINRELEF-LD also provides indices at the bivariate level and on an item-by-item basis.

We shall describe now the bivariate indicators. As discussed below, the input of the programs requires the user to specify the pairs of items that are locally dependent and also the residual correlations corresponding to each pair. With this information, the program can obtain the indices: I_{LD}, I_{LI}, and RE_{LD} for each specified doublet. This information is expected to be highly relevant for item selection purposes, as illustrated in the example below.

At the item-by-item level, finally, SINRELEF-LD provides, for each item, the familiar reliability estimate that would be expected if this item was deleted. It will also provide the expected relative change in information if this item was deleted. To fix ideas, we shall return to our numerical illustration above. Recall that, for the raw scores based on the entire item set ω_{LD}=0.88 and I_{LD}=7.33. Suppose now that, if a given item, say j, was deleted, the reliability estimate would be expected to drop to $\omega_{LD\text{-}j}$=0.86. In information terms, the drop would then be: $I_{LD\text{-}j}$=6.14. Finally, the relative change of information would be: (6.14-7.33)/7.33=-0.16, i.e., other things equal, an information loss of 16% would be expected if this item was deleted. Now, can we afford to remove this item?. The reader familiar with statistics will likely respond that we must first determine if these differences are significant, and she would be right. So, to address this point, SINRELEF-LD provides confidence intervals for the reported estimates. Again, the reader is referred to Ferrando et al. (2024) for technical details.

So far, we have summarized all the indices the program provides, and have organized them at three levels: total-score, bivariate, and item-by item. However, SINRELEF-LD is even more comprehensive, and considers to alternative modeling approaches from which the items may have been calibrated. The first modeling is linear FA, which, in calibration terms, means that the items have been calibrated by assuming that their scores are (approximately) continuous and unbounded. The second modeling is non-linear factor analysis based on an Underlying Variables Approach (UVA). In calibration terms, this second choice means that the item scores are treated as ordered-categorical variables. For a thorough discussion of this topic, which (a) is relatively non-technical and (b) focuses on the item calibration process, the reader is referred to Ferrando & Lorenzo-Seva (2014).

In closing, Table 1 summarizes the features implemented in SINRELEF-LD

Table 1

Features provided by SINRELEF-LD at each different level

	Feature	Overall scores	Bivariate	Item-by-item
Local dependency	Information	I_{LD}	I_{LD}	$I_{LD\text{-}j}$
	Reliability	ω_{LD}		$\omega_{LD\text{-}j}$
	Relative efficiency	RE_{LD}	RE_{LD}	RE_{LD}
Local independency	Information	I_{LI}		
	Reliability	ω_{LI}		

2. Installation and setup

The proposal so far discussed has been implemented in an R package called SINRELEF.LD (Information, Reliability, & Relative efficiency under Local Dependences). SINRELEF.LD has been developed in R Version 4.0.2 and runs with R versions more recent than 3.5.0. R program is available through their website (https://cran.r-project.org), and it has versions for Windows, macOS and Linux distributions.

SINRELEF.LD package can be downloaded manually from CRAN repository at:

https://cran.r-project.org/package=SINRELEF.LD

Alternatively, the package can be downloaded using R console using the following syntax:

```
> install.packages("SINRELEF.LD")
```

Once the package is installed and loaded, the main function is available for the user, which will be described in the next section.

3. Program Usage 1: Entering and importing data

The program uses as input the calibration item estimates obtained from fitting extended unidimensional FA-solutions, in which the modeled local dependences are included. All the implemented procedures can be obtained from (a) linear FA solutions in which the items are treated as approximately continuous or (b) non-linear solutions in which the item scores are treated as ordered-categorical.

The inputs required for SINRELEF.LD are described as:

L: A vector containing the item loading estimates in the unidimensional factor-analytic (FA) solution.

PSI: A vector containing the item residual standard deviations.

THRES: A vector containing the item thresholds. (for the graded model only).

ncat: Number of item response categories (for the graded model only).

model: The factor-analytic model used for calibrating the item set, being 'linear' for the standard FA model or 'graded' for the non-linear ordered-categorical FA model.

doublet_list: A vector containing the pairs of items with correlated specificities.

cor_doublet: A vector containing the residual correlations corresponding to each pair, specified in doublet_list.

N: The number of observations in the sample.

CI: Choice of the confidence level for the intervals (90 or 95, 90 by default).

display : Determines if the output will be displayed in the console, TRUE by default. If it is TRUE, the output is returned silently and if it is FALSE, the output is returned in the console.

SINRELEF.LD was designed with three possible outputs in mind: Mplus, FACTOR and lavaan R package. Then, the inputs required for the function are fairly similar to the corresponding outputs in those statistical programs.

Most precisely, for Mplus users, the procedure for obtaining the input data for SINRELEF.LD assessment is described as the following:

For linear model, the user requires: a) the loadings vector from the estimate column of the model results; b) The residual variances of each variable, also found in model results and c) The residual correlations, which must be obtained from the “with” elements in the standardized model results.

For the graded model, all the structural estimates must be obtained from the standardized model results. Then, the user requires: a) the loading vector from the standardized model results; b) the residual correlations obtained from the “with” elements in the standardized model results and c) the thresholds of each variable as shown in the output (the first column on the thresholds section, as a column vector).

For FACTOR users, the process for obtaining the input data for SINRELEF.LD is very similar to the process of obtaining from an Mplus output. The main difference is that, since FACTOR work with correlation structures, continuous and graded solutions are standardized. Then, there is no need to provide residual variances in both cases. Another difference is the formatting of the thresholds in the graded case. The user can provide the b values as a matrix, as presented in the FACTOR output, or re-arrange them for obtaining a single vector, like the Mplus scenario.

Finally, for lavaan output, the user has to extract from the lavaan fitted model the same required inputs as previously described. For linear model: a) the loadings vector, b) residual variances and c) residual correlations. For the graded model: a) the loadings vector, b) residual correlations and c) the thresholds.

4. Program Usage II: Syntax and output

The R package includes only one main function, also called SINRELEF.LD, where the user has to provide the required inputs in order to implement the procedures aforementioned.

```
SINRELEF.LD(L, PSI, THRES, ncat, model = 'linear', doublet_list, cor_doublet, N, CI = 90, display = TRUE)
```

The function returns a R list, including all the output variables from the analysis. This list includes:

omld: Omega reliability estimate taking in to account the item local dependences.

omli: Omega reliability estimate if all the items were locally independent.

relef: Score relative efficiency.

relef_doublet: Doublet score relative efficiency.

omega_del: Reliability estimate if the item was omitted.

r_info_del: Relative information if the item was omitted.

5. Illustrative Example

For this example, 400 subjects were randomly selected from those who participated in the study by Dueñas et al. (2022), which focused on the involvement of parents in the education of their sons and daughters. The participants were therefore parents with teenage children enrolled in high-school. These parents completed the Family Involvement Questionnaire-High School Version (FIQ-HS), but for this example we have only used their data in the Home-based activities subscale. The 17 items of this subscale, which have a Likert-type response format (rarely, sometimes, often and always), focus on parental activities outside of school that promote learning, such as talking to teenagers about careers and schooling and helping them with homework. We have chosen this subscale because it contains some redundant items, with a very similar wording or content (Dueñas et al., 2022). In fact, previous analyses already showed that the error terms of four pairs of items were substantially correlated.

were either very similarly worded or tapped similar content.

A unidimensional solution, based on the nonlinear UVA FA model, and in which the four doublets referred to above were freely estimated, was fitted to this data by using robust ULS estimation as implemented in Mplus software. Goodness-of-fit results were acceptable: RMSEA and 90% C.I.= 0.057 (0.049;0.067); CFI=0.91; GFI=0.95.

The obtained loading vector (`LAM`) was the following:

Item	Loading
Y1	.367
Y2	.507
Y3	.646
Y4	.590
Y5	.679
Y6	.714
Y7	.619
Y8	.670
Y9	.538

Y10	.745
Y11	.560
Y12	.372
Y13	.356
Y14	.549
Y15	.709
Y16	.583
Y17	.631

The pairs of items with correlated specificities (doublets) were: 3-10, 12-13, 14-16 and 11-15; these should be provided as a numerical vector, containing all the pairs of items: c(3, 10, 12, 13, 14, 16).

The correlations between doublets were:

Pair of items	r
Y3 – Y10	.666
Y12 – Y13	.687
Y14 – Y16	.439
Y11 – Y15	.333

These correlations also should be provided as a numeric vector: c(.666, .687, .439, .333).

Finally, the thresholds (`thres`) should be provided as obtained on Mplus matrix (vertical vector):

Thresholds	Loading
Y1$1	-0.845
Y1$2	-0.313
Y1$3	0.388
Y1$4	0.923
Y2$1	-0.904
Y2$2	-0.298

Y2$3	0.353
Y2$4	0.946
Y3$1	-0.962
Y3$2	-0.350
Y3$3	0.303
Y3$4	0.912
Y4$1	-0.834
Y4$2	-0.230
Y4$3	0.434
Y4$4	1.024
Y5$1	-0.923
Y5$2	-0.372
Y5$3	0.380
Y5$4	0.958
Y6$1	-0.935
Y6$2	-0.340
Y6$3	0.342
Y6$4	0.970
Y7$1	-0.915
Y7$2	-0.335
Y7$3	0.364
Y7$4	0.904
Y8$1	-1.049
Y8$2	-0.443
Y8$3	0.353
Y8$4	0.915
Y9$1	-1.089
Y9$2	-0.385
Y9$3	0.337

Y9$4	1.049
Y10$1	-0.974
Y10$2	-0.321
Y10$3	0.364
Y10$4	0.871

The thresholds should also be provided as a numeric vector. c(-.845, -.313, .388, …)

With this information, the input for running the program is:

```
SINRELEF.LD(L = LAM, THRES = thres, ncat = 4, model = 'graded', doublet_list = c(3,10,12,13,14,16), cor_doublet = c(.666,.687,.439,.333), N = 400, CI = 90, display = TRUE)
```

We shall now discuss a series of steps, based on the results provided by SINRELEF-LD, that will allow the user to identify redundant items, and to determine which items could be removed if necessary.

The first step is to inspect the standardized loadings estimates already provided above in the Lavaan output. As can be seen, the most poorly functioning items, with the lowest standardized loadings, are 1,12, and 13, while the items with the highest estimates are 6,10, and 15.

The second step is the comparison between the Omega reliability estimate in which the local dependencies are considered (Omega-LD), and the "ceiling" estimate if the items were locally independent (Omega-LI). As can be seen in Table 3, Omega-LI is higher than Omega-LD. In fact, only Omega-LD significantly differs from .80, which is the value usually considered as a minimum threshold (e.g. Raykov & Marcoulides, 2011), as indicated by the fact that its confidence interval does not include this value. Furthermore, the Score Relative efficiency suggest that there is a 19% loss of information/efficiency due to the modeled local dependences in the scale.

Table 2

Omega reliability estimates under local independence and under local dependency, and score Relative efficiency

Omega-LI		Omega-LD		Score Relative efficiency	
Value	90% CI	Value	90% CI	Value	90% CI
.80	[.78, .82]	.76	[.74, .79]	.81	[.77, .85]

The third step is inspecting the four specified salient doublets, whose Relative efficiencies are shown in Table 4. It should be considered that Doublet Relative efficiency values close to .50 indicate high redundancy, which means that neither item provides additional information beyond that provided by the other. In contrast, values closer to 1 would involve very little or no redundancy. According to the results, the more prominent doublets are the pairs 3-10 and 12-13, with relative efficiency values closer to .50. Therefore, these are the two doublets that will be taken into account in the next step, described below.

Table 4

Doublet Relative efficiencies and 90% confidence intervals

Doublet	RE	90% CI
3-10	.63	[.52, .74]
12-13	.58	[.47, .70]
14-16	.67	[.61, .73]
11-15	.76	[.70, .81]

Finally, the fourth step is inspecting the Omega-LD reliability estimates, and the corresponding Relative information change, after removing one item at a time. In the case of scales with redundant items, the results of this step, in conjunction with the previous ones, makes it easier the decision of which items could be eliminated to reduce redundancy. Regarding the most prominent doublet (pair 12-13), both Omega-LD and Relative information change are the same if item 12 or item 13 are removed, as can be seen in Table 5. As both items (a) contribute with low amounts of information (both have low standardized loadings), and (b) are highly redundant, removing one or the other results in equivalent Relative information change, and there does not seem to be a compelling rationale for selecting one or the other as a candidate for deletion. Regarding the other salient doublet (pair 3-10), removing item 3 or 10 also results in similar

Omega-LD and relative information changes. However, as item 10 has a higher item relative efficiency, it seems preferable to remove item 3.

Table 5

Omega reliability estimates under local dependency after deleting each item, and the Relative information change

Item	Omega-LD without the item	Relative information change
1	.76	-.04
2	.75	-.06
3	.77	.02
4	.74	-.10
5	.75	-.09
6	.76	-.02
7	.75	-.06
8	.76	-.02
9	.75	-.08
10	.77	.01
11	.75	-.07
12	.78	.07
13	.78	.07
14	.76	-.03
15	.75	-.09
16	.75	-.05
17	.74	-.10

Considering these results, we decided to remove items 3 and 13. Table 6 shows the estimated results for the total scores of the remaining 15 items. The Omega-LD is again lower than the Omega-LI, but now the confidence intervals of both values include the .80 value. The Score Relative efficiency is higher than the one previously obtained for the 17 items. Therefore, removing items 3 and 13 has led to a reduction of the loss of information/efficiency due to the modeled local dependences in the scale

Table 6

Omega reliability estimates under local independence and under local dependency, and score Relative efficiency, after removing items 3 and 13

Omega-LI		Omega-LD		Score Relative efficiency	
Value	90% CI	Value	90% CI	Value	90% CI
.80	[.78, .81]	.78	[.77, .80]	.90	[.89, .92]

References

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 903-913.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

DeMars, C. E. (2020). Comparing Causes of Dependency: Shared Latent Trait or Dependence on Observed Response. *Journal of Applied Measurement*, *21*(4), 400-419.

Dueñas, J. M., Morales-Vives, F., Camarero-Figuerola, M., Tierno-García, J. M. (2022). Spanish adaptation of The Family Involvement Questionnaire-High School: Version for parents. *Psicología Educativa*, *28*(1), 31-38. https://doi.org/10.5093/psed2020a21

Ferrando, P. J., & Lorenzo-Seva, U. (2014). Exploratory item factor analysis: Additional considerations. *Anales de Psicología*, *30*(3), 1170-1175.

Ferrando, P. J., & Morales-Vives, F. (2023). Is it quality, is it redundancy, or is model inadequacy? Some strategies for judging the appropriateness of high-discrimination items. *Anales de Psicología*, *39*(3), 517-527. http://dx.doi.org/10.6018/analesps.535781

Ferrando, P. J., Hernández-Dorado, A., & Lorenzo-Seva, U. (2023; in press). A Simple Two-Step Procedure for Fitting Fully Unrestricted Exploratory Factor Analytic Solutions with Correlated Residuals. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-9. https://doi.org/10.1080/10705511.2023.2267181

Ferrando, P. J., Navarro-González, D., & Morales-Vives, F. (2024; submitted). Linear and Nonlinear Indices of Score Accuracy and Item Effectiveness for Measures that Contain Locally Dependent Items.

McDonald, R. P. (1982). Linear vs. nonlinear models in Item Response Theory. *Applied Psychological Measurement*, *6*, 379-396. https://doi.org/10.1177/014662168200600402

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423