Psicológica (2015), 36, 143-163.

Assessing retest effects at the individual level: A general IRT-based approach

Pere J. Ferrando*

'Rovira i Virgili' University, Spain

Test-retest studies for assessing stability and change are widely used in different domains and allow improved or additional individual estimates of interest to be obtained. However, if these estimates are to be validly interpreted the responses given at Time-2 must be free of retest effects, and the fulfilment of this assumption must be empirically checked. This article proposes a comprehensive item response theory-based approach for assessing retest effects at the individual level and test the assumption of local independence under repetition. The approach can be used with a wide array of unidimensional and multidimensional models, and is based on correlation-type and mean-square-type indices. Procedures for (a) establishing critical values for detection purposes and (b) interpreting the magnitude of the retest effects for the detected respondents are also proposed. Furthermore, the article discusses the consequences of not addressing retest effects in stability and change studies. The procedures were assessed with simulation and used in three empirical studies. In all cases they worked well and provided meaningful information.

Test-retest (T-R) designs with a short-term retest interval are quite common in personality and attitude measurement. They are widely used in: (a) reliability assessment (APA/AERA/NCME, 1999, Morrison, 1981), (b) the clinical assessment of trait changes due to the effects of treatments (Finkelman, Weiss, & Kim-Kang, 2010, Reise & Haviland, 2005), (c) the measurement of attitude change (Sherif, Sherif & Nebergall, 1981), and (d) personnel selection for gauging the effects of test-coaching and practice (Hausknecht, Trevor & Farr, 2002).

* Acknowledgments: The research was supported by a grant from the Spanish Ministry of Economy and Competitiveness (PSI2011-22683). Correspondence should be sent to: Pere Joan Ferrando. Research Centre for Behavioral Assessment. Universidad 'Rovira i Virgili'. Facultad de Psicología. Carretera Valls s/n. 43007 Tarragona (Spain). Tel: +34 977 558079. E-mail: perejoan.ferrando@urv.cat

In the applications mentioned above, the most common T-R design is a multiple-indicator design in which a test made up of multiple items is administered with an interval shorter than two months (Cattell, 1986). In a stability-based application (e.g. a reliability analysis) no treatment is given during the retest interval and the conditions of administration are the same at both points in time. So, the trait level of each respondent is assumed to be the same at Time-1 and Time-2 (Goldberg, 1963). In a change-based application, the trait level is assumed to be different at Time-1 and Time-2, because a treatment is given between Time-1 and Time-2 or because the conditions of administration at the two points in time are different.

As discussed below, in both stability and change applications, the information gained from the repeated administration of the test can be used to obtain improved or additional individual estimates of interest. These estimates, however, can only be validly interpreted if the assumptions on which the analysis is based are met. In particular, the item response theory (IRT) modelling of most T-R designs assumes (implicitly or explicitly) that the local independence principle holds for the repeated measurements. More specifically, for a fixed trait level, the conditional distributions of the responses to the same item in two repeated administrations are assumed to be independent of each other (Nowakowska, 1983). This assumption is denoted here as local independence under repetition (LIR).

In a short-term design the LIR assumption will not be met if retest effects (REs) are operating. In this article "retest effects" are defined as the tendency for individuals to duplicate their former item responses, either because they recall them or because of incidental item features which tend to elicit the same response on each occasion (APA/AERA/NCME, 1999, Morrison, 1981). REs defined in this way are a case of "positive surface local dependence" (Chen & Thissen, 1997, Houts & Edwards, 2013). More complex types of REs might be envisaged (e.g. Arendasy & Sommer, 2013). For example, memory effects in a change study might lead, in some cases, to more differentiated responses. These scenarios, however, will not be considered here.

So far, assessment of REs has been addressed either at the test-score level by using a descriptive approach, or at the item level by using a structural equation modelling approach in which REs are modelled via correlated residuals (e.g. Ferrando, 2001). This article, however, will take a different approach. First, REs will be addressed from an IRT framework and treated as a particular case of local dependence. Second, they will be assessed not at the test or the item level, but at the level of each individual respondent (see Ferrando, 2010, 2014). Overall, the aim is to propose a

comprehensive IRT-based approach for addressing REs at the individual level that can be used with a variety of response formats and under a wide range of IRT models.

The rest of the article is organized as follows. First, a background of concepts and preliminary results is provided. Second, the proposed procedures are presented. Third, the impact of REs at the individual level is discussed, thus justifying the interest of the present proposal. Finally, the functioning of the procedures is assessed and illustrated by means of empirical studies.

General Background

Consider a test of n items that is administered to the same respondents at two points in time with a given retest interval. Let X_{ij} be the response of individual i to item j at Time 1, and X'_{ij} the corresponding response at Time 2. The responses can be binary (scored as 0 and 1), graded (scored by successive integers) or continuous, and the response patterns at both points of time are assumed to be well fitted by the same unidimensional or multidimensional IRT model. The item parameter estimates are assumed to be fixed and known (see e.g. Zimowski, et al., 2003) so, in the scoring stage, individual trait estimates are obtained on the basis of the fixed item parameters.

The information that is gained from the repeated administration of the items can be used in the scoring stage to obtain more accurate trait estimates (in stability studies) or additional individual estimates (in change studies). In a stability study, a common trait estimate can be obtained by treating the test-retest pattern as if it were a single response pattern made up of $2n$ responses. Because the estimate is now obtained from a pattern that is twice as long, it is expected to have less measurement error than that obtained on the sole basis of the Time-1 data (see Ferrando, 2014). In a change study, two individual trait estimates are obtained from the separate response patterns, and their difference serves as a basis for assessing individual change (e.g. Finkelman, Weiss, & Kim-Kang, 2010, Reise & Haviland, 2005).

Preliminary Results

This proposal is based on two conditional expectations for which I shall provide general results. The first expectation, denoted by $E(X_j|\theta)$ is the expected item score for fixed θ. The second, denoted by $\sigma^2(X_j|\theta)$ is the conditional variance for fixed θ. In general θ will be vector-valued, and will reduce to a scalar in the case of a unidimensional model.

In the binary case the conditional expectations are

$$E(X_j \mid \boldsymbol{\theta}_i) = P_j(\boldsymbol{\theta}_i)$$
$$\sigma^2(X_j \mid \boldsymbol{\theta}_i) = P_j(\boldsymbol{\theta}_i)(1 - P_j(\boldsymbol{\theta}_i)) \quad (1)$$

where $P_j(\boldsymbol{\theta})$ is the conditional probability of scoring 1 on item j.

In the graded response case, they are given by (Chang & Mazzeo, 1994).

$$E(X_j \mid \boldsymbol{\theta}_i) = \sum_r r P_{jr}(\boldsymbol{\theta}_i)$$
$$\sigma^2(X_j \mid \boldsymbol{\theta}_i) = \left[\sum_r r^2 P_{jr}(\boldsymbol{\theta}_i)\right] - \left[E(X_j \mid \boldsymbol{\theta}_i)\right]^2 \quad (2)$$

where $P_{jr}(\boldsymbol{\theta})$ is the conditional probability of scoring in category r (r=1,2..) in item j.

Finally, continuous responses are usually fitted with the linear factor-analytic model (FA). If this is the case, the conditional expectations are

$$E(X_j \mid \boldsymbol{\theta}_i) = \mu_j + \sum_k \lambda_{jk}\theta_{ik}$$
$$\sigma^2(X_j \mid \boldsymbol{\theta}_i) = \sigma^2_{ej} \quad (3)$$

where μ_j is the item intercept, λ_{jk} is the loading of item j on factor k, and σ^2_{ej} is the residual variance of item j.

Assessing REs at the Individual Level: The Present Proposal

The basic indices proposed in this section are of two types – correlation-based and mean-square-based – and can be used with all the item formats and types of IRT model discussed above.

Correlation-based indices: rti Q3

Yen (1984) proposed Q3 as an index for quantifying local dependence between a pair of binary items. Q3 was defined as the product-moment

correlation between the residuals of item *j* and item *k* computed over the sample of respondents. Ferrando (2014) proposed using the Q3 rationale at the individual level instead of the item level, and defined an index that measures the degree of local dependence under repetition for a given respondent. This article proposes a more general formulation of the index so that it can be used with the family of IRT models discussed in the sections above.

Define first the residual scores as

$$\delta_{ij} = X_{ij} - E(X_j \mid \boldsymbol{\theta}_i) \quad ; \quad \delta'_{ij} = X'_{ij} - E(X'_j \mid \boldsymbol{\theta}'_i), \tag{4}$$

Consider now the $n\times1$ vector $\boldsymbol{\delta}_i$ containing the residual scores of respondent *i* for the *n* items at Time-1, and let $\boldsymbol{\delta}'_i$ be the corresponding vector at Time-2. The index *rti Q3* (where *rti* is used to refer to retest effects at the individual level) is defined as the product-moment correlation between $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}'_i$: $rti\ Q3_i = \rho(\boldsymbol{\delta}_i, \boldsymbol{\delta}'_i)$.

If the IRT model is correct, and LIR holds for all the responses of individual *i*, then δ_{ij} and δ'_{ij} are all random error scores and the expected value of *rti Q3* for this individual is zero. On the other hand, if REs are operating, the terms δ_{ij} and δ'_{ij} are positively correlated. So, the expected value of *rti Q3* is positive, and increases with the strength of the REs. Finally, it should be noted that the index has been defined by using the 'true' trait values of the individual which are unknown. Therefore, *rti Q3* is always computed using trait estimates.

Mean-square residual-based indices: rti MSR and rti MSRW

Let $D_{ij}=X_{ij}-X'_{ij}$ be the difference between the score of respondent *i* to item *j* at Time-1 and his/her score to the same item at Time-2. The conditional expectation and variance of D_{ij} are given by:

$$E(D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\theta}'_i) = E(X_{ij} \mid \boldsymbol{\theta}_i) - E(X'_{ij} \mid \boldsymbol{\theta}'_i). \tag{5}$$

And

$$\sigma^2(D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\theta}'_i) = \sigma^2(X_{ij} \mid \boldsymbol{\theta}_i) + \sigma^2(X'_{ij} \mid \boldsymbol{\theta}'_i) - 2Cov(X_{ij}, X'_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\theta}'_i). \tag{6}$$

If LIR holds, then the covariance term in (6) vanishes. If it does, the scaled statistic

$$z(D_{ij}) = \frac{D_{ij} - E(D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)}{\sqrt{\sigma^2(X_{ij} \mid \boldsymbol{\theta}_i) + \sigma^2(X'_{ij} \mid \boldsymbol{\theta}'_i)}} \tag{7}$$

is conditionally distributed as a standard variable with zero expectation and unit variance.

The *rti MSR* index is now defined as

$$rti\ MSR(i) = \sum_{j=1}^{n} \frac{z^2(D_{ij})}{n}. \tag{8}$$

So, *rti MSR* is a mean square statistic which can range from 0 to infinity and which has an expectation of 1 if LIR holds. Assume now, however, that REs are operating. If they are, the covariance term in (6) does not vanish but has a positive value. So, the 'true' variance term in (6) is smaller than that assumed in the denominator of (7) and, therefore, the expected value of *rti MSR* is smaller than 1. So, low values of *rti MSR* are indicative of REs. Finally, as with *rti Q3*, the index is defined by using the 'true' trait levels but in practice is always computed using trait estimates.

Mathematically, expression (8) has the same form as the mean-square outfit statistics which are used for assessing fit in Rasch analysis (e.g. Smith, Schumacker & Bush, 1998). This similarity allows some well known problems to be anticipated for *rti MSR*. In particular, the statistic is expected to be very sensitive to outliers when the conditional variances in the denominator of (7) are small. In order to diminish the potential effect of outliers, a weighted version of *rti MSR,* which is termed *rti MSRW*, is now proposed. Mathematically, *rti MSRW* has the same form as the infit statistics used in Rasch analysis (e.g. Smith et al., 1998).

$$rti\ MSRW(i) = \frac{\sum_{j=1}^{n}[\sigma^2(X_{ij}|\boldsymbol{\theta}_i) + \sigma^2(X'_{ij}|\boldsymbol{\theta}'_i)]z^2(D_{ij})}{\sum_{j=1}^{n}[\sigma^2(X_{ij}|\boldsymbol{\theta}_i) + \sigma^2(X'_{ij}|\boldsymbol{\theta}'_i)]} = \frac{\sum_{j=1}^{n}[D_{ij} - E(D_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)]^2}{\sum_{j=1}^{n}[\sigma^2(X_{ij}|\boldsymbol{\theta}_i) + \sigma^2(X'_{ij}|\boldsymbol{\theta}'_i)]} \quad (9)$$

Like its unweighted counterpart, *rti MSRW* ranges from 0 to infinity, has an expectation of 1 if LIR holds, and the expectation goes towards 0 as the impact of REs increases. The main difference with (8) is that now each squared standardized residual is weighted by the conditional variance so that the influence of the less informative responses, which are the potential outliers, can be reduced.

Assessing index accuracy, determining critical values, and interpreting the indices

If the indices proposed above are to be used for detection purposes, then reference distributions must be determined under the null hypothesis for establishing cutpoints or critical values. In principle, theoretical distributions could be considered for both types of index. On the one hand, *rti Q3* is a product-moment correlation, so a Fisher-r-to-z transform is expected to bring its values close to the normal distribution. On the other hand, *rti MSR* and *rti MSRW* are expected to approach a scaled chi-square distribution as the number of responses increases. Experience with related indices at the item level, however, suggests that in both cases these approximations do not closely adhere to the reference distribution (Chen & Thissen, 1997, Houts & Edwards, 2013). So, the approach proposed here is to obtain the reference distributions and associated cutpoints by simulating the distribution of the index of interest for each $\boldsymbol{\theta}$ value (see e.g. van Krimpen-Stoop & Meijer, 2002). In more detail, the proposal proceeds in three steps: (a) for fixed item parameters and the trait estimate/s of the individual, simulate a large number of item response patterns under the LIR assumption (i.e. the null hypothesis), (b) compute the index from each

pattern and obtain its empirical distribution, and (c) determine the critical value by computing the desired percentile of the empirical distribution.

Once it has been detected that a response pattern has potentially been impacted by REs, the values of the indices can also serve to assess the magnitude of these effects. For this purpose, *rti Q3* is possibly the clearest index. It is a standard product-moment correlation expected to provide positive values when REs are operating, and so its range of values of interest is between 0 and 1. For their part, *rti MSR* and *rti MSRW* are discrepancy measures that do not have an upper bound. However, their range of interest is between 1 and 0 (i.e. the lower tail). Values lower than 1 mean that the discrepancies between the scores at Time 1 and the corresponding scores at Time 2 are smaller than those expected given the stochastic nature of the IRT model. This tendency to repeat the responses beyond what the model is able to predict is what is interpreted as REs.

Interpretation of both types of index would be enhanced if a confidence interval were also provided in addition to the point estimate, and the approach proposed here is to determine these intervals by using nonparametric Bootstrap. The standard percentile-method seems to be quite appropriate here because it is very simple, tends to produce stable interval lengths, and is suitable for small samples. For mean-squared statistics and for product-moment correlations, however, it typically produces poor coverage accuracy with respect to the nominal level (Schenker, 1985, Sievers, 1996). So, the proposed approach is to obtain confidence intervals by using the double percentile Bootstrap (e.g. Hall & Martin, 1988).

Relevance of the Proposal: The Impact of REs on the Individual Estimates

Consider first the common trait estimate obtained under the stability assumption in the unidimensional case. If the item parameters are fixed and known, then the presence of REs is not expected to produce biases in this estimate. However, it is expected to reduce its accuracy. Conceptually, both results are clear. The tendency to repeat the responses is not expected to change in any systematic way the trait estimate that would be obtained if only the responses given at Time-1 are considered. At the same time, however, the responses at Time-2 are redundant to a greater or lesser extent, and so they are expected to provide less information than could be obtained if additional new items were used. To see the second result more formally, consider that, if the repeated responses are locally dependent, then the test information estimated by the sum of item informations will overstate the 'true' amount of information. So, when REs are operating (a) the point

estimate of the trait level will be less accurate, and (b) the standard error of this estimate will be incorrect. Result (b) implies that both the confidence intervals around the trait estimates and the test statistics for assessing trait differences are likely to be misleading. These distortions, in turn, might have consequences in both individual assessment and validity studies.

I turn now to the change scenario in which I shall consider the Z-test proposed by Finkelman, Weiss and Kim-Kang (2010) as the basic procedure for estimating individual change. Define first the change point estimate δ as $\delta_i = \hat{\theta}_i' - \hat{\theta}_i$. If the item parameters are assumed to be the same at Time-1 and Time-2, then the Z-test can be written as

$$|Z| = \frac{|\delta_i|}{\sqrt{2\frac{1}{I(\hat{\theta}_{i\,pool})}}} \qquad (10)$$

where $\hat{\theta}_{pool}$ is the pooled trait estimate based on the full *2n* pattern. In other words, it is the common trait estimate proposed here under the stability assumption (i.e. the best estimate under the null hypothesis of no change).

The standard error in the denominator of (10) is based on the amount of information corresponding to the common trait estimate, so the results are those discussed above: If REs are operating, the test information estimate is upwardly biased, so the standard error estimate in (10) is incorrect and downwardly biased.

If LIR holds, then each expected score at Time-2 depends solely on *θ'*. However, if REs are operating, then the expected score also depends on the response which was given at Time-1, and, in the limiting case of perfect local dependence, the expected score at Time-2 depends only on (and is the same as) the score at Time-1. Overall, then, the presence of REs brings the Time-2 scores closer to the Time-1 scores than the model predicts. So, the expected trait estimate at Time-2 is biased towards the estimate at Time-1, and, therefore, |*δ*| is biased towards zero (i.e. attenuated). To sum up, the presence of REs is expected to (a) decrease the point estimate of the amount of individual change, and (b) make the corresponding standard error incorrect.

As in the stability case, results (a) and (b) above might have a negative impact on both individual assessment and validity assessment. However, the impact is expected to be more serious here: results (a) and (b) imply that change is not expected to be well detected in those respondents for whom REs

are strong. This distortion, in turn, might have important consequences for individual decisions taken in settings such as clinical treatment or attitude-change assessment.

EMPIRICAL STUDIES

The functioning of the indices and procedures proposed in this paper was assessed with two initial simulations and three real-data studies in the personality domain. In all the studies only the main results are provided. Further details can be obtained from the author.

Two Initial Simulation Studies

Given the preliminary nature of the studies, most basic conditions were kept as simple as possible. First, the simulations were based on the stability scenario. Second, only the least parameterized models were considered. So, the first simulation was based on binary responses that behaved according to the two-parameter model (2PM) while the second was based on continuous responses that behaved according to the unidimensional linear FA model. In both studies the fixed conditions were as follows: two 'pseudo samples' of 600 cases each were generated from the corresponding model by using item discriminations normally distributed between 0.2 and 0.8, and thresholds/intercepts normally distributed between -2.0 and 2.0. As for the scoring, the distribution of θ was standard normal, and the trait estimates were Bartlett factor scores (continuous case) and Bayes EAP scores (binary case). EAP estimates were chosen in the second simulation in order to avoid potential problems of non-covergence or implausible estimates, especially in the short-test conditions. Finally, in the first pseudo sample (null hypothesis conditions) no REs were operating in any of the simulees. In the second (alternative hypothesis conditions), REs were operating for all simulees.

The common independent variables were test length and magnitude of the REs. Test lengths were n=20 (short test), n=40 (medium test) and n=60 (long test). The levels in the REs variable (alternative hypothesis conditions) were determined by the value of the correlation between the measurement errors, and were r=0.20 (weak REs), r=0.40 (medium REs), and r=0.60 (strong REs). In all cases, the critical values for determining significance were based on 500 simulated response patterns, and the values which are reported in each cell are the average obtained across 30 replications.

In the null-hypothesis simulations the dependent variables were: (a) mean value, and (b) empirical type-I error using a unilateral contrast (i.e. upper tail for *rti* Q3 and lower tail for *rti MSR* and *rti MSRW*) for nominal

levels of 0.10 and 0.05. In the second case, they were: (a) mean value and (b) proportion of hits (individual with REs detected as such) when the 0.10 and 0.05 type-I nominal levels were used. The results are shown in tables 1 (binary responses) and 2 (continuous responses).

Table 1. Results of Simulation Study 1: Binary Variables.

	n=20					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	.99	.00	.98	.00	.98
F.D. prop (α= 0.10)	.10	.11	.10	.10	.10	.11
F.D. prop (α= 0.05)	.06	.06	.05	.06	.05	.06
REs subsample						
Mean	.15	.82	.28	.69	.44	.54
Hit prop (α= 0.10)	.26	.33	.39	.46	.59	.65
Hit prop (α= 0.05)	.15	.19	.27	.32	.45	.48

	N=40					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	.99	.00	.99	.00	.99
F.D. prop (α= 0.10)	.10	.11	.10	.10	.10	.11
F.D. prop (α= 0.05)	.05	.06	.05	.05	.06	.06
REs subsample						
Mean	.16	.82	.30	.69	.45	.54
Hit prop (α= 0.10)	.32	.37	.57	.58	.82	.82
Hit prop (α= 0.05)	.22	.24	.43	.43	.70	.70

	n=60					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	.99	.00	.98	.00	.98
F.D. prop (α= 0.10)	.10	.11	.10	.11	.10	.11
F.D. prop (α= 0.05)	.06	.06	.06	.06	.06	.07
REs subsample						
Mean	.15	.83	.30	.70	.45	.54
Hit prop (α= 0.10)	.39	.41	.66	.64	.89	.88
Hit prop (α= 0.05)	.27	.28	.52	.50	.82	.81

Table 2. Results of Simulation Study 2: Continuous Variables.

	n=20					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	1.00	.00	1.00	.00	1.00
F.D. prop (α= 0.10)	.10	.10	.10	.11	.10	.11
F.D. prop (α= 0.05)	.05	.05	.05	.06	.05	.06
REs subsample						
Mean	.20	.80	.39	.60	.59	.40
Hit prop (α= 0.10)	.34	.27	.68	.59	.93	.92
Hit prop (α= 0.05)	.23	.17	.54	.52	.87	.86

	n=40					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	1.00	.00	1.00	.00	1.00
F.D. prop (α= 0.10)	.10	.10	.10	.10	.10	.11
F.D. prop (α= 0.05)	.05	.05	.05	.05	.05	.05
REs subsample						
Mean	.20	.80	.39	.60	.59	.40
Hit prop (α= 0.10)	.48	.38	.89	.81	.99	.99
Hit prop (α= 0.05)	.35	.25	.80	.69	.99	.99

	n=60					
	ree'= .20		ree'= .40		ree'= .60	
	rti Q3	rtiMSRW	rti Q3	rti-MSRW	rti Q3	rtiMSRW
No REs subsample						
Mean	.00	1.00	.00	1.00	.00	1.00
F.D. prop (α= 0.10)	.10	.10	.10	.11	.10	.10
F.D. prop (α= 0.05)	.05	.05	.05	.05	.05	.05
REs subsample						
Mean	.20	.80	.40	.60	.60	.40
Hit prop (α= 0.10)	.57	.46	.96	.91	1.00	1.00
Hit prop (α= 0.05)	.45	.32	.92	.85	1.00	1.00

Summary of results

In all cases it was found that *rti MSRW* systematically outperformed *rti MSR*. So only the *rti MSRW* results are presented here.

Overall, the results behaved as expected. Under the null-hypothesis conditions the mean values were generally close to the expectations and the empirical type-I values agreed with the nominal levels. Under the alternative-hypothesis conditions, the means of *rti Q3* and *rti MSRW* departed from the null expected values in the predicted direction, and the departures were more pronounced as the REs increased. As for detection

power, in all cases the proportion of hits increased with test length and magnitude of REs, as it should.

Regarding comparisons, two results are worth discussing. First, both indices tended to work better in the case of continuous responses, which is to be expected given that product-moment correlations and mean squares perform generally better with continuous variables. Second, in the continuous case *rti Q3* tended to perform better than *rti MSRW* under the alternative-hypothesis conditions. However, in the binary case, *rti MSRW* showed more power than *rti Q3* when the test was short and the REs were small. Finally, the differences between both indices gradually decreased as test length and amount of REs increased.

As a summary, the general results suggest that both *rti Q3* and *rti MSRW,* would be good at detecting individuals with strong REs in tests of 40 or more items. However, this result must be qualified mainly for two reasons. First, the simulation assumes that the chosen IRT model exactly holds in the population of simulees. Second, the item parameters are assumed to be known. These conditions, indeed, are never met in practice but are expected to be well approximated if the item parameters have been calibrated in a large and independent sample and model-data fit is good. In less favorable conditions, such as when minor factors not accounted for by the model are impacting the responses or when the item estimates have large sampling variability, a loss of power of the proposed statistics can be expected (at the very least). So, it is safe to say that further simulation based on more realistic scenarios is strongly needed.

Real-Data Example 1. A Test-Retest Study Based on Binary Responses

The first example is based on a data set used in Ferrando, Lorenzo-Seva and Molina (2001). A 60-item questionnaire for measuring Neuroticism (N) was administered on two occasions to a sample of university students under the same conditions and with a 4-week retest interval. The responses were binary and the items were calibrated by fitting the 2PM to the Time-1 sample data (N=625). As detailed in the original study, the model-data fit was acceptable.

In the present analysis, EAP scores were estimated for the 432 respondents who were present at both administrations. Because the study was based on stability assumptions, the scores were obtained by treating the 120 responses as if they formed a single pattern. Finally, the *rti Q3* and *rti MSRW* estimates were obtained based on the calibrated item parameters and the EAP scores. The means were: 0.47 (*rti Q3*) and 0.53 (*rti MSRW*). The

product-moment correlation between *rti Q3* and *rti MSRW* was r=-0.91, which suggests strong agreement between both indices.

In both cases the means discussed above suggest that strong REs were operating for most of the respondents. To provide more support for this result, simulated data based on the calibrated item parameters but in which the LIR assumption was met was generated for 500 simulees. In agreement with the expectations, the means of the simulated data were: 0.00 (*rti Q3*) and 1.00 (*rti MSRW*). Figure 1 shows the distribution of the simulated *rti MSRW* values (dashed line) together with the distribution of the real values (solid line). Note that the distribution of the simulated *rti MSRW* values is centered at 1 and clearly shifted to the right.

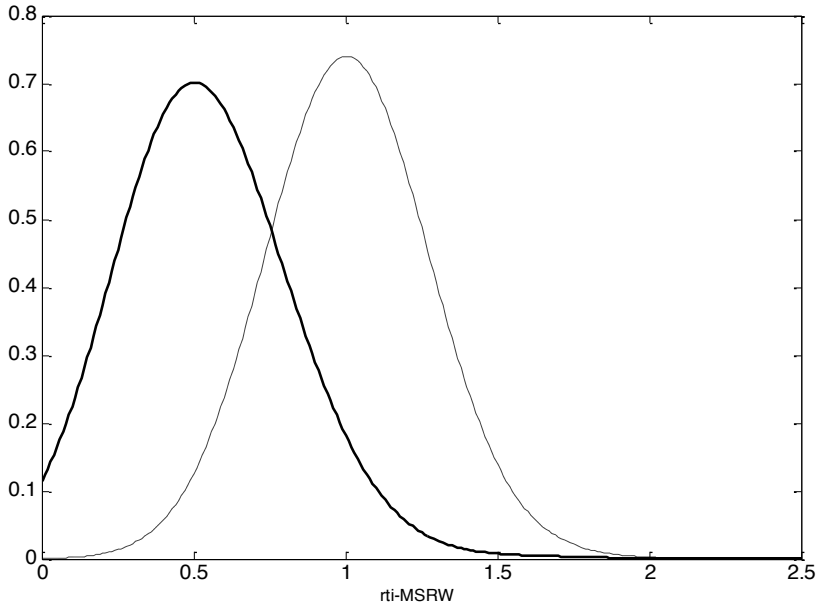

Figure 1. Distribution of real (solid line) and simulated (dashed line) *rti-MSRW* values. Illustrative example 1.

In the last step, cutpoint values were determined by using 500 replications per individual, and a 90% confidence level. According to the *rti Q3* results, 92% of the respondents would be flagged as potentially impacted by REs. According to *rti MSRW* the corresponding percentage would be 88%.

As a final illustration, table 3 summarises the results of two respondents with similar trait estimates but possibly affected very differently by REs. For each index, the reported results are the point estimate and the Double Bootstrap 90% confidence interval obtained by using 2000 replications at the first level and 44 at the second level.

Table 3. Summary of results for respondents 55 and 201. Illustrative example 1.

Respondent nº	rti-Q3	90% C.I.	rti-MSRW	90% C.I.	$\hat{\theta}$
55	-.03	(-.24;.20)	1.51	(1.25;1.79)	-.39
201	.85	(.70;.95)	.16	(.02;.30)	-.41

It was detected that participant 201 was impacted by REs, and inspection of the estimates and confidence intervals suggests that this impact was rather strong. In contrast, the results suggest that the scores of participant 55 were not affected by REs in the slightest. Even though the trait estimates of both respondents are quite similar, the estimate of respondent 55 is expected to be far more accurate than that of respondent 201 because the 'true' information is based on a response pattern made up of 120 responses which are locally independent. In contrast, there is much less information for respondent 201, and it is probably largely the same as that would be obtained by using solely the Time-1 data (i.e. 60 items).

Real-Data Example 2. A Study Based on Binary Responses and Change Assumptions

The second example uses a subtest made up of 21 items taken from the 60-item scale discussed in the previous example, and was administered on two occasions with a retest interval of 6 weeks in a group of university students. Of these, 332 completed the questionnaire at Time 1, and 277 participated in both administrations. At Time-1 the participants were asked to respond under standard instructions. At Time-2 they were given faking-inducing instructions that were expected to produce a substantial decrease in the N scores. So, unlike the first example, a pretest-treatment-post-test design was used in which a temporal change in the trait levels was assumed.

As in the previous example, the 21 items were calibrated according to the 2PM by using the Time-1 data. The fit was considered to be acceptable: the goodness of fit index estimate was GFI=0.93 and the standardized root mean squared residual was 0.06. So, EAP individual scores were next obtained. In this example, however, the EAP estimates were computed separately at Time-1 and Time-2 based on the common item parameters. The mean trait estimates together with the 90% confidence intervals were: 0.00 (-0.07; 0.07) at Time-1 and -1.37 (-1.46:-1.28) at Time-2. Finally, the means of *rti Q3* and *rti MSRW* values were: 0.20 (*rti Q3*) and 1.01 (*rti MSRW*). The product-moment correlation between both indices was r=-0.87.

Taken together, the results just discussed clearly suggest that (a) at the group level there is a marked general decrease in the N levels at Time-2 (as expected), and (b) unlike the previous example, the impact of REs in this case is small or negligible for most of the respondents. These results, however, are still compatible with non-informative or non-valid estimates of faking-induced change for certain individuals.

Figure 2 shows the $|\delta|$ individual estimates of change (see equation 10) plotted against the corresponding *rti Q3* values. The scatterplot is markedly heteroscedastic, and is a clear example of Fisher's (1959) "twisted pear" effect, in which the change estimates become increasingly variable as the impact of REs (measured by *rti Q3*) decreases. When REs are not present (*rti Q3* values around 0) the change estimate is not restricted in any way. However, as discussed above, when REs are strong (*rti Q3* values approaching 1) the change estimates are necessarily small. So, for the individuals in the bottom right-hand square of the graph, the estimates of change cannot be validly interpreted because they are likely to be grossly attenuated.

Real-Data Example 3. A Study Based on the Multiple FA model and Stability Assumptions

The Overall Personality Assessment Scale (OPERAS; Vigil_Colet et al., 2013) is a multidimensional questionnaire intended to measure the Big-Five personality dimensions. It consists of 5 short scales of 7 items each that measure the content dimensions, plus a 4-item control scale intended to measure social desirability, and the response format for all of the items is 5-point Likert. The OPERAS is particularly suitable for illustrating the use of the multidimensional versions of the indices proposed here. If the unidimensional indices had been used on a separate scale by scale basis, the response patterns would have been too short for achieving accurate and

powerful detection of REs. However, the use of the multidimensional versions obtains the indices on the basis of a 39-response pattern, which should be enough for them to function appropriately provided that the sources of REs generalize across subtests. On the other hand, if the sources were (at least in part) scale specific, the multidimensional versions might be insensitive to overall patterns that show REs on few specific subscales.

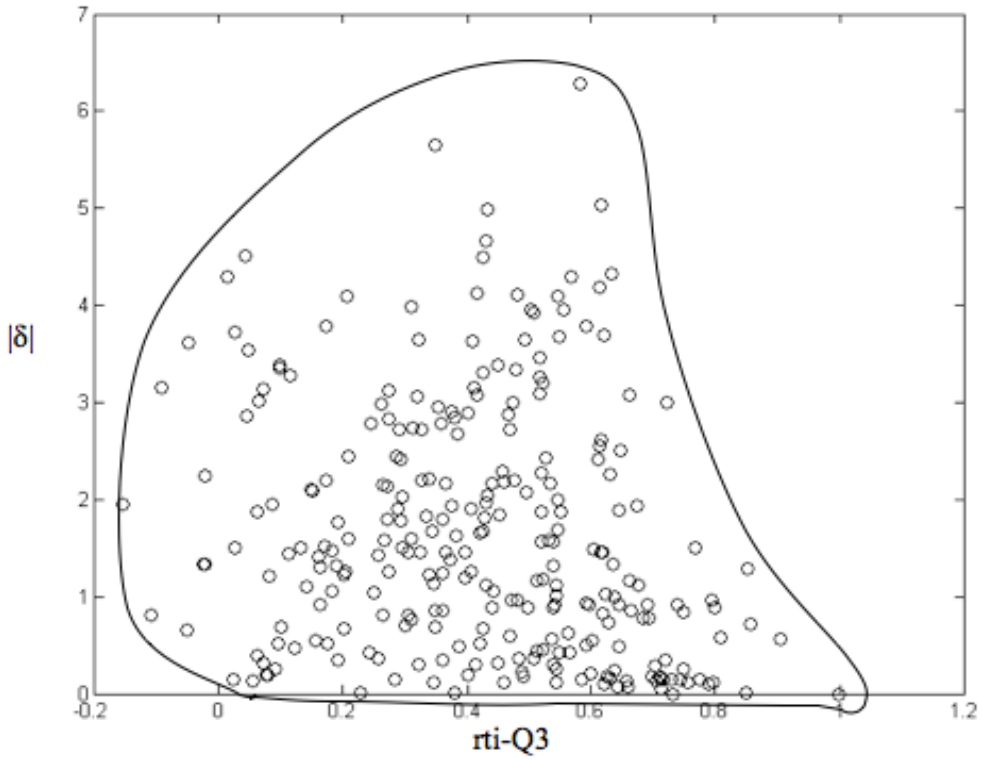

Figure 2. Scatterplot of trait change estimates against *rti-Q3* values. Illustrative example 2.

The present example is based on a sample of 128 respondents that was collected to assess test-retest reliability by using a retest interval of 3 weeks. Because this sample is relatively small, the item parameter estimates that were used were those obtained from the normative sample of 3,838 respondents (Vigil-Colet et al., 2013). The model used in the calibration was unrestricted linear FA in five orthogonal dimensions and the solution

was a clear independent-clusters basis which matched the a priori allocation of the items.

In the scoring stage, Bartlett's factor scores were obtained in each of the five dimensions for the 128 respondents and, given that the study used stability assumptions, the scores were pooled estimates obtained by treating the 78 responses as if they formed a single pattern. Finally, the multidimensional continuous versions of *rti Q3* and *rti MSRW* were computed for each respondent. The means were 0.72 (*rti Q3*) and 0.63 (*rti MSRW*), and suggest that strong REs were operating for most of the respondents. The percentages of potentially impacted respondents were 88% (*rti Q3*) and 70% (*rti MSRW*). So, the results in this case suggest that the sources of REs at the individual level generalize across subtests. Finally, the product-moment correlation between *rti Q3* and *rti MSRW* was r=-0.87 which, again, indicates close agreement between both indices.

DISCUSSION

REs can adversely affect the individual estimates obtained in T-R studies thus making valid interpretations of these estimates questionable. Therefore, it would be advisable in a study of this type to systematically use the procedures proposed here with each participant before his/her estimates are interpreted or used for purposes of assessment or in further validity studies.

The procedures proposed in this article can be used in both stability and change studies under a variety of response formats and IRT models. So, overall, this is a wide-ranging proposal that cannot be considered as totally finished, and some issues clearly must be developed further. I discuss below two of these issues.

First, the indices and approaches proposed here are expected to work well in the favorable conditions discussed above: a large and independent calibration sample, a long test, and good model-data-fit results. However, as also discussed above, how they are expected to work under less favorable conditions is still largely unknown. This point requires further intensive research under a variety of potentially relevant conditions to be fully explored.

Second, the replication-based approaches proposed here for making inferences about the indices seem to work well but are computationally very intensive, particularly the double Bootstrap for obtaining confidence intervals. Although, as discussed above, strict adherence to normal and chi-square distributions cannot be expected, these approximations might be

enough for most practical purposes and, if so, the inference processes would be greatly simplified. Whether this is or not the case requires again intensive simulations under realistic conditions as well as further studies based on real data.

In spite of the acknowledged limitations, however, the preliminary results obtained are clearly encouraging. The simulation results suggest that the indices behave as expected and that, under favorable conditions, REs can be detected with tests of moderate length especially if these effects are moderate to strong (the most relevant condition in practice). As for the empirical examples, they provided meaningful results in all cases, and these results should be taken into account in applied research. The results of studies 1 and 3 suggest that in stability designs, relatively short retest intervals of 3 or 4 weeks still give rise to strong REs for most of the respondents. In contrast, study 2 was based on a change design with a longer retest interval, and in this case the REs were far weaker. However, the scatterplot in figure 2 suggests that valid change estimates cannot be obtained for the individuals who were impacted by REs.

In closing, I would like to discuss a practical and clearly needed further development. The procedures proposed here are relatively simple and can be easily programmed but if the present proposal is to be put to widespread use a free user-friendly program must be available. This is a clear aim for the future.

RESUMEN

Evaluación de los efectos retest a nivel individual: Un enfoque global basado en la TRI. Los estudios test-retest para evaluar estabilidad y cambio se utilizan ampliamente en diferentes dominios y permiten obtener estimaciones individuales más precisas o información adicional sobre el individuo. Sin embargo, para interpretar de forma válida las estimaciones derivadas de un estudio de este tipo, las respuestas obtenidas en la segunda ocasión deben estar libres de efectos retest, y el cumplimiento de este requisito debe evaluarse empíricamente. El presente artículo presenta un enfoque general basado en la teoría de respuesta al ítem para evaluar los efectos retest a nivel del individuo y evaluar el supuesto de independencia local bajo repetición. El enfoque propuesto puede utilizarse con una amplia variedad de modelos unidimensionales y mutidimensionales, y está basado en índices correlacionales e índices mínimo-cuadráticos. Se proponen procedimientos para (a) establecer valores críticos con fines de detección y (b) interpretar la magnitud de los efectos retest para los individuos detectados. Se discuten además las consecuencias de ignorar los efectos retest cuando están actuando. Los procedimientos propuestos se evaluaron mediante simulación y se aplicaron en tres estudios empíricos. En todos los casos funcionaron bien y proporcionaron información de interés.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Arendasy, M.E. & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*, 181-192.

Chang, H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response function in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Chen.W. & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics,* 22, 265-289.

Cattell, R. B. (1986). The psychometric properties of tests: Consistency, validity and efficiency. In R. B. Cattell & R. C. Johnson (Eds.), *Functional psychological testing* (pp. 54-78). New York: Brunner/Mazel.

Ferrando, P.J.; Lorenzo-Seva, U & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement. 25*, 3-19.

Ferrando, P.J. (2001). An IRT-based two-wave model for studying short-term stability in personality measurement. *Applied Psychological Measurement*, *26*, 286-301.

Ferrando, P.J. (2010). Assessing short-term individual consistency using IRT-based statistics. *Psicológica* , *31*, 319-334.

Ferrando, P.J. (2014). A Comprehensive approach for assessing Person Fit with Test–Retest data. *Educational and Psychological Measurement* (in press; available on line doi:10.1177/001316441351855

Finkelman, M.D., Weiss, D.J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34,* 238-254.

Fisher, J. (1959). The twisted pear and the prediction of behavior. *Journal of Consulting Psychology, 23,* 238-254.

Goldberg, L.R. (1963). A model of item ambiguity in personality assessment. *Educational and Psychological Measurement*, *23*, 467-492.

Hall, P. & Martin, M.A. (1988). On Bootstrap Resampling and Iteration. *Biometrika*, *75*, 661-671.

Hausknecht, J.P., Trevor, C.O., & Farrr, J.L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87*, 243-254.

Houts, C.R. & Edwards, M.C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37,* 541-562.

Morrison, D.G. (1981). A stochastic model for test-retest correlations. *Psychometrika, 46,* 143-151.

Nowakowska, M. (1983). *Quantitative psichology: some chosen problems and new ideas*. Amsterdam: North-Holland.

Reise, S.P., & Haviland, M.G. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84,* 228-238.

Schennker, N. (1985). Qualms about Bootstrap confidence intervals. *Journal of the American Statistical Association, 80*, 360-361.

Sievers, W. (1996). Standard and Bootstrap confidence intervals for the correlation coefficient. *British Journal of Mathematical and Statistical Psychology, 49*, 381-396.

Smith, R. M., Schumacker, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Sherif, C.W., Sherif, M., & Nebergall, R.E. (1981). *Attitudes and Attitude Change*. Westport: Greenwood Press.

van Krimpen-Stoop, E.M.L.A. & Meijer, R.R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*, 164-180.

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., Lorenzo-Seva, U. (2013). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema, 25*, 100-106.

Yen, W.M. (1993). Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software.

(Manuscript received: 29 May 2014; accepted: 8 September 2014)