Psicológica (2015), 36, 391-417.

A combined IRT and SEM approach for individual-level assessment in test-retest studies

Pere J. Ferrando[*]

'Rovira i Virgili' University, Spain

The standard two-wave multiple-indicator model (2WMIM) commonly used to analyze test-retest data provides information at both the group and item level. Furthermore, when applied to binary and graded item responses, it is related to well-known item response theory (IRT) models. In this article the IRT-2WMIM relations are used to obtain additional information for each individual respondent. Procedures are proposed for (a) obtaining individual estimates of trait levels and amount of change, and (b) assessing whether the main assumptions on which these estimates are based are met. The procedures are organized in a comprehensive approach that can be used with binary, graded, and continuous responses. The relevance of the proposal is discussed and guidelines are given on how to use the approach in applied research. Finally, the approach is illustrated with an empirical data set. It worked well and provided meaningful information.

Test-retest (T-R) studies in which a test made up of multiple items is administered twice to the same respondents at an interval of less than two months are very common in psychological measurement and, particularly, in personality and attitude measurement (e.g. Cattell, 1986). Studies of this type are used, for example, in the assessment of (a) trait changes due to the effects of clinical treatments (Finkelman, Weiss, & Kim-Kang, 2010, Reise & Haviland, 2005, Weiss & Von Minden, 2011), (b) attitude change (Aish & Jöreskog, 1990), and (c) effects of test-coaching and practice in personnel selection (Hausknecht, Trevor & Farr, 2002). At the methodological level, the two main existing approaches for analyzing the common type of T-R

[*] The research was supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2014 SGR 73) and by a grant from the Spanish Ministry of Education and Science (PSI2014-52884-P). Correspondence should be sent to: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicología. Carretera Valls s/n. 43007 Tarragona (Spain). Tel: +34 977 558079. E-mail: perejoan.ferrando@urv.cat

design used in these studies are: (a) Item Response Theory (IRT)-based analysis (e.g. Fischer, 2003, Grimm, Kuhl & Zhang, 2013, Reise & Haviland, 2005, Wang & Wu, 2004,Weiss & Kinsbury, 1984), and (b) analysis based on a structural equation model (SEM) with latent variables (e.g. Aish & Jöreskog, 1990, Ferrando, 2002, Kenny & Campbell, 1989).

The standard SEM for analyzing the T-R design considered here is the simple and well known two-wave multiple-indicator model (2WMIM; Jöreskog, 1979), which: (a) is intended for continuous indicators, and (b) provides information which focuses exclusively on the item and group level. The model, however, can be extended in directions that have been much less exploited. First, when applied to discrete item responses by using categorical-variable methodology, the 2WMIM becomes a longitudinal extension of well-known IRT models, Second, the model can be extended so that it can also be used for individual assessment. Regarding this latter point, it is noted that in many applications, particularly when posttest change is to be assessed, information at the individual level can be as relevant as that obtained at the group level (Kruyen, Emons & Sijtsma, 2013, Weiss & Von Minden, 2011).

The present article exploits the IRT-2WMIM relations and proposes procedures that allow the researcher to make a detailed assessment at the individual level on the basis of the results provided by the 2WMIM. More specifically, this article develops a comprehensive approach for (a) testing whether the main assumptions on which the model is based are fulfilled at the level of each individual respondent, and, if so, (b) using the item and group information provided by the 2WMIM for obtaining more accurate trait levels and change estimates for a given individual.

SEM-based proposals have already been made about the individual assessment of both change and appropriateness, but appear to be only indirectly related to those made in this article. As for change, longitudinal SEMs at the individual level have been proposed in the context of growth modeling (e.g. Andrade & Tavares, 2005, Embretson, 1991, Grimm, Kuhl & Zhang, 2013). However, growth models assess a trajectory of changes as a function of time and are based on multiple measurement occasions. SEM-based procedures have also been proposed for assessing individual model appropriateness, but generally with the aim of improving overall model-data fit (e.g. Bollen & Arminger, 1991, Raykov & Penev, 2002). On the other hand, the proposals that are more directly concerned with individual assessment (Ferrando, 2007, Reise & Widaman, 1999) do not consider T-R-based models.

Review of the 2WMIM and Results

Consider a test of *n* items that is administered to the same *N* respondents at two points in time with a short retest interval. Let $X^{(1)}_{ij}$ be the score of individual *i* to item *j* at Time 1, and $X^{(2)}_{ij}$ the corresponding score at Time 2. I shall describe the application of the 2WMIM to three types of item formats: (a) binary (scored as 0 and 1), (b) graded in *c* points (scored by successive integers and treated as ordered categorical variables) and (c) graded or more continuous treated as continuous-unlimited variables. The common framework I use is Muthén's (1984) underlying variables approach, and can be summarized as follows. In cases (a) and (b) the observed scores are assumed to arise as a result of a discretization of a latent response variable (Y_j) governed by a single threshold τ_j in the binary case, and by *c-1* thresholds in the graded case. Thus, at Time *m* (*m*=1 or 2) the relation in the binary case is

$$\begin{aligned} X^{(m)}{}_{ij} &= 0 \; if \; Y_{ij}^{(m)} < \tau^{(m)}{}_j \\ X^{(m)}{}_{ij} &= 1 \; if \; Y_{ij}^{(m)} \geq \tau^{(m)}{}_j \end{aligned} \quad . \qquad (1)$$

And in the graded-response case it is

$$\begin{aligned} &X^{(m)}{}_{ij} = 1 \quad if \quad Y_{ij}^{(m)} < \tau^{(m)}{}_{j1} \\ &X^{(m)}{}_{ij} = 2 \quad if \quad \tau^{(m)}{}_{j1} \leq Y_{ij}^{(m)} < \tau^{(m)}{}_{j2} \\ &\ldots\ldots. \\ &X^{(m)}{}_{ij} = c \quad if \quad \tau^{(m)}{}_{jc-1} \leq Y_{ij}^{(m)} \end{aligned} \quad . \qquad (2)$$

Finally, in the continuous case Y_j is assumed to be identical to the observed score (i.e. $Y^{(m)}_{ij}= X^{(m)}_{ij}$). In cases (1) and (2) the *Y's* cannot be observed and so identification constraints are needed. The initial constraints considered here are that the *Y's* in (1) and (2) are normally distributed and that their means and variances at Time-1 are zero and one, respectively.

The 2WMIM is intended for the Y_j indicators, and so the model is the same in the three cases above. It consists of: (a) two measurement submodels (one at Time 1 and one at Time 2) and (b) a latent variable submodel relating the trait measured at Time-1 to the trait measured at Time-2. The main practical differences due to the different types of score are that in cases (a) and (b) the model is fitted to the item threshold and the inter-item tetrachoric/polychoric matrices, whereas in the continuous case it is fitted to the mean vector and the inter-item covariance matrix. Overall then, the structural modeling considered here is two-stage: the mean/threshold vectors and covariance/correlation matrices are obtained from the data observed in the first stage, and then the 2WMIM is fitted to these vectors and matrices in the second stage. With regards to the first stage, in the categorical cases the tetrachoric/polychoric correlations are assumed that are estimated by using a two-step procedure in which thresholds are estimated in the first step without any equality restrictions, and correlations are then estimated in the second step for given thresholds. As for the second stage, given that the latent submodel described below is saturated, the measurement submodels and the latent variable submodel are assumed to be fitted simultaneously in this stage.

The measurement submodel at Time *m* is Spearman's single-factor model

$$Y_{ij}^{(m)} = \mu_j^{(m)} + \lambda_j^{(m)}\,\theta_{im} + \varepsilon_{ij}^{(m)} \tag{3}$$

where μ_j is the intercept, λ_j is the factor loading, and the residuals ε_{ij}*'s* have zero means and are uncorrelated with θ and among themselves. For fixed θ, Y_j is normally distributed with zero mean and residual variance $\sigma_{\varepsilon j}^{2\,(m)}$.

With appropriate reparameterization, when submodel (3) is applied to case (1) it becomes the IRT two-parameter normal-ogive model. Applied to case (2), it becomes Samejima's (1969) normal-ogive graded response model (see, e.g., Ferrando, 2002). For the sake of simplicity, however, I shall only use the FA parameterization here.

The latent variable submodel relating the trait at Time-1 to the trait at Time-2 is

$$\theta_{i2} = \alpha + \beta\theta_{i1} + \zeta_i . \tag{4}$$

And states that the trait at Time-2 is a linear additive function of the trait at Time-1 with intercept α, slope β, and a random disturbance ζ, which is assumed to be uncorrelated with θ_1 (Kenny & Campbell, 1989). θ_m can be identified in different ways (see Little, Slegers & Card, 2006). The one proposed here is to set the mean and variance of θ_1 to 0 and 1, respectively, and, relative to this scaling, freely estimate the mean and variance of $\theta_{2.}$ These constraints, together with the measurement invariance constraints discussed below, are sufficient to identify the model.

According to the latent identification constraints discussed above, the mean and variance of θ_2 are estimated as

$$\begin{aligned} E(\theta_2) &= \alpha \\ Var(\theta_2) &= \beta^2 + Var(\zeta) \end{aligned} \tag{5}$$

The coefficient of stability is then defined as the correlation between the trait levels at Time-1 and the trait levels at Time-2 (Jöreskog, 1979). It is given by

$$\rho_{\theta 1,\theta 2} = \frac{Cov(\theta_1,\theta_2)}{\sqrt{Var(\theta_1)}\sqrt{Var(\theta_2)}} = \frac{\beta}{\sqrt{Var(\theta_2)}}. \qquad (6)$$

It is of special relevance in studies on individual change. A low value of $\rho_{\theta 1,\theta 2}$ essentially means that the rank order of the individuals as far as their trait levels is concerned has changed substantially between Time-1 and Time-2. This result, in turns, means that change is not (approximately) constant for all individuals, but affects individuals differently.

I shall now discuss two basic points that must be considered when specifying the complete 2WMIM. First, when the same items are measured at two points in time, the literature consistently recommends that the errors corresponding to the same indicators should be correlated (e.g. Jöreskog, 1979, Kenny & Campbell, 1989, Pitts, West & Tein, 1996). This is because part of the measurement error in this case may not be random, but systematic, thus making the responses locally dependent under repetition (e.g. Ferrando, 2015). This lack of local independence may be due to memory effects, incidental item features which tend to elicit the same response on each occasion, or even external influences, among others (Ferrando, 2015, Pitts, West & Tein, 1996). In this article I shall (a) refer generically to local dependence under repetition as "retest effects" (REs; Ferrando, 2014, 2015), (b) assume that the dependence is positive (a point that can be empirically assessed), and (c) not try to differentiate further sub-components or sources of local dependence. So, overall, the residuals between the responses to the same item on both occasions ($\varepsilon^{(1)}_j$ and $\varepsilon^{(2)}_j$) are allowed to covary and this covariance is expected to be positive.

As for the second point, (a) the clarity in the interpretation of the stability/change processes and (b) the accuracy and stability of the model's parameter estimates increase as the degree of measurement invariance in the measurement submodels increases (Little, 2013, Millsap & Meredith, 2007, Grimm, Kuhl & Zhang, 2013). The ideal condition, then, would be that of strict invariance (Millsap & Meredith, 2007), in which the item thresholds/intercepts, loadings and residual variances are the same at Time-1 and Time-2. Experience, however, suggests that this condition is generally unrealistic in practice, and so unnecessarily restrictive (see Ferrando, 2002, and Little, 2013, for a discussion). On the other hand, the strong invariance condition in which the thresholds/intercepts and loadings are constrained but the residual variances are not can be reasonably attained in many T-R studies, and allows for a clear interpretation of stability and change both at the group and the individual level. Finally, it should be stressed that strong invariance together with the identification constraints discussed above are more than enough to identify the 2WMIM. In fact, the model could be identified with only one threshold and one loading constrained to be invariant over time. So, to sum up, the model with strong invariance is clear, parsimonious, over-identified and likely to lead to accurate and stable estimates. However, the present proposal can also be applied to less constrained solutions in which some of the items are not invariant over time.

When the 2WMIM is fitted to a given dataset, the results provide three main pieces of information: (a) model appropriateness, (b) item properties, and (c) structural group-level results. Model appropriateness is a basic requisite for interpreting results (b) and (c), and is assessed at the entire group level by conducting a standard model-data fit investigation for the full model and/or for the component submodels (e.g. Aish & Jöreskog, 1990).

Item-level information has three parts. First, there is the information regarding the quality of the items as measures of θ, which is obtained via intercepts/thresholds, loadings, and residual variances. Second, there is the information about the degree of invariance over time as discussed below. Finally, information about the magnitude of the REs is obtained by inspecting the item residual covariances or correlations discussed above.

At the group level the 2WMIM provides two main pieces of information. The first is the degree of trait stability between the first and second administration, which is estimated from the coefficient of stability. The second is the information about the group mean trait levels and group variances at Time-1 (fixed to 0 and 1) and at Time-2 (freely estimated), which allow the amount of change at the group level to be assessed. By noting that (a) the trait mean at Time-2 is directly a measure of mean group change, and (b) the standard error of the mean at Time-2 is also provided, the significance of group change can be assessed by using a confidence interval with the form

$$\bar{\theta}_2 \pm z_c se(\bar{\theta}_2) \tag{7}$$

where z_c is the value in the standard normal distribution that cuts off the desired percent of cases in the middle of the distribution. If group change is found to be significant, a Cohen's-*d*-type effect-size measure (Cohen, 1988) for assessing the practical magnitude of the change can be obtained as

$$d(\bar{\theta}_2) = \frac{\bar{\theta}_2}{\sqrt{Var(\theta_1 - \theta_2)}} \tag{8}$$

where

$$Var(\theta_1 - \theta_2) = 1 + Var(\theta_2) - 2\rho(\theta_1, \theta_2)\sqrt{Var(\theta_2)} . \tag{9}$$

The rest of this section will discuss the results that are needed to develop the procedures proposed. I shall first use the generic expression $P(X_j/\theta)$ to denote the conditional probability (discrete case) or conditional density (continuous case) assigned to a specific item score for fixed θ. For the models considered in this article, this general expression leads to the following results:

(a) binary case

$$P(X^{(m)}{}_{ij} = 1 \mid \theta_m) = \int_{-\infty}^{\frac{\lambda^{(m)}{}_j\theta_m - \tau^{(m)}{}_j}{\sigma^{(m)}_{\varepsilon j}}} \phi(z)dz = \Phi\left(\frac{\lambda^{(m)}{}_j\theta_m - \tau^{(m)}{}_j}{\sigma^{(m)}_{\varepsilon j}}\right) \tag{10}$$

where $\phi(u)$ is the density function of a standard normal variable.

(b) graded-response case

$$P(X^{(m)}{}_{ij} = r \mid \theta_m) = \Phi\left(\frac{\lambda^{(m)}{}_j\theta_m - \tau^{(m)}{}_{jr-1}}{\sigma^{(m)}_{\varepsilon j}}\right) - \Phi\left(\frac{\lambda^{(m)}{}_j\theta_m - \tau^{(m)}{}_{jr}}{\sigma^{(m)}_{\varepsilon j}}\right). \tag{11}$$

For r=1,...c, and

(c) continuous case

$$P(X^{(m)}{}_{ij} \mid \theta_m) = \phi\left(\frac{X^{(m)}{}_{ij} - \mu^{(m)}_j - \lambda^{(m)}{}_j\theta_m}{\sigma^{(m)}_{\varepsilon j}}\right) \tag{12}$$

Next, I shall use the general terms $E(X_j|\theta)$ and $\sigma^2(X_j|\theta)$ to denote the expected item score and its conditional variance fixed θ. Again, the general expressions give rise to the following results:

(a) binary case

$$\begin{aligned} &E(X^{(m)}{}_{ij} \mid \theta_m) = P(X^{(m)}{}_{ij} = 1 \mid \theta_m) \\ &\sigma^2(X^{(m)}{}_{ij} \mid \theta_m) = P(X^{(m)}{}_{ij} = 1 \mid \theta_m)(1 - P(X^{(m)}{}_{ij} = 1 \mid \theta_m)) \end{aligned} \tag{13}$$

(b) graded-response case

$$E(X^{(m)}{}_{ij} \mid \theta_m) = \sum_r rP(X^{(m)}{}_{ij} = r \mid \theta_m)$$
$$\sigma^2(X^{(m)}{}_{ij} \mid \theta_m) = \left[\sum_r r^2 P(X^{(m)}{}_{ij} = r \mid \theta_m)\right] - \left[E(X^{(m)}{}_{ij} \mid \theta_m)\right]^2 \qquad (14)$$

(c) continuous case

$$E(X^{(m)}{}_{ij} \mid \theta_m) = \mu^{(m)}{}_j + \lambda^{(m)}{}_j \theta_m$$
$$\sigma^2(X^{(m)}{}_{ij} \mid \theta_m) = \sigma^{(m)2}_{\varepsilon j} \qquad (15)$$

Estimation of Change at the Individual Level

Let us assume that the 2WMIM has been fitted and that the model-data fit is acceptable. The measurement and structural estimates provided by the model will next be taken as fixed and known and used to obtain estimates of change for (a) each individual in the group, or (b) new individuals belonging to the population in which the SEM holds.

Let $\mathbf{x}_i$ be the full vector of responses given by individual *i* at Time-1 and Time-2 and $\boldsymbol{\theta}_i = [\theta_{i1,}\ \theta_{i2}]$ the 'true' trait levels of this individual. The likelihood of $\mathbf{x}_i$ for any of the three types of responses considered in the article can be written generically as

$$L(\mathbf{x_i} \mid \boldsymbol{\theta}_i) = \prod_{j=1}^{n} P(X^{(1)}_{ij} \mid \boldsymbol{\theta}_i) \prod_{j=1}^{n} P(X^{(2)}_{ij} \mid \boldsymbol{\theta}_i). \qquad (16)$$

Specific expressions of (16) for the three response formats considered here are provided in the appendix.

Maximum likelihood (ML) estimates of $\boldsymbol{\theta}_i$ are the pair of values that maximize (16). Now, ML estimation uses only the information provided by the measurement sub-models but ignores the information provided by the structural part of the model. In order to make full use of the SEM-based information I propose to use Bayes expected a posteriori (EAP; Bock & Mislevy 1982) estimation. The EAP estimate of $\boldsymbol{\theta}_i$, is the mean of the posterior distribution of $\boldsymbol{\theta}_i$ given $\mathbf{x}_i$

$$EAP(\boldsymbol{\theta}_i) = \int_\theta \boldsymbol{\theta} L(\mathbf{x_i} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \qquad (17)$$

The term *g(θ)* in (17) is the joint bivariate density of **θ** and contains the information provided by the structural part of the 2WMIM. I propose to set *g(θ)* as bivariate normal (but other specifications are possible), with the mean and variance of θ_1 equal to 0 and 1, the mean and variance of θ_2 as the corresponding structural estimates, and $\rho(\theta_1,\theta_2)$ as the coefficient of stability in (6). Estimation of (17) is standard, and, in practice, the integral is approximated as accurately as required using numerical quadrature.

The posterior covariance matrix is

$$\sum(\boldsymbol{\theta} \mid \mathbf{x}_\mathbf{i}) = \int_\theta (\boldsymbol{\theta} - EAP(\boldsymbol{\theta}_i))^2 L(\mathbf{x}_\mathbf{i} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$= \begin{bmatrix} PSD^2(\theta_{i1}) & PCv(\theta_{i1},\theta_{i2}) \\ PCv(\theta_{i1},\theta_{i2}) & PSD^2(\theta_{i2}) \end{bmatrix}. \qquad (18)$$

As the number of items increases, the distribution of the EAP estimates in (17) approaches normality (Chang & Stout, 1993), and the posterior standard deviations (PSDs) in (18) become equivalent to asymptotic standard errors (Bock & Mislevy, 1982).

Conceptually, the EAP individual estimates proposed so far are IRT trait estimates. However, unlike the IRT estimates in common use, those proposed here do make full use of the SEM-based information discussed above, and so they are expected to be more accurate.

Let $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ be the EAP point estimates corresponding to individual *i*. The model-based estimate of change for this individual is now defined as

$$\hat{\delta}_i = \hat{\theta}_{i2} - \hat{\theta}_{i1} \qquad (19)$$

with (posterior) standard error (see equation 18)

$$PSD(\hat{\delta}_i) = \sqrt{PSD^2(\hat{\theta}_{i2}) + PSD^2(\hat{\theta}_{i1}) - 2PCv(\hat{\theta}_{i1},\hat{\theta}_{i2})} \qquad (20)$$

Given the normality-approximation result discussed above, for a test of reasonable length it is proposed to assess $\hat{\delta}_i$ statistically by using a normal-based confidence interval approach (strictly speaking, a credibility interval). This interval, which is the individual-level counterpart to the group-level interval in (7), can be constructed as

$$\hat{\delta}_i \pm z_c PSD(\hat{\delta}_i) \qquad (21)$$

Significance of change can be assessed by inspecting whether the zero value falls outside the interval. If the amount of change is considered to be significant, a relative, scaled measure for assessing the magnitude of this change can be obtained as

$$d(\hat{\delta}_i) = \frac{\hat{\delta}_i}{\sqrt{Var(\theta_1 - \theta_2)}}. \qquad (22)$$

Index (22) can be regarded as the individual counterpart of the effect-size measure (8). It assesses the magnitude of individual change in standard deviation units with respect to the distribution of the trait differences in the group.

The point estimate of change (19) and the confidence-interval approach in (21) are IRT-based conditional measures of change that are population independent. Therefore, they avoid the unreliability problems associated with conventional gain scores (see Mellenbergh, 1999). Essentially, they can be regarded as a refinement of a series of proposals which were made in the context of IRT (e.g. Fischer, 2003, Kruyen, Emons & Sijtsma, 2013, Reise & Haviland, 2005, Weiss & Kinsbury, 1984). The present proposals, however, (a) are based on a full SEM whose appropriateness is rigorously tested, and (b) use most of the information that can be obtained from the 2WMIM estimates. In contrast, IRT-based procedures (a) generally use the calibration results obtained from a pretest sample or the results obtained solely from the Time-1 data, (b) ignore the structural part of the model, and (c) do not test the different possible invariance conditions but generally assume directly strict item invariance of the item parameters (see e.g. Ferrando, 2014) or use approximate parameter-drift procedures (Wang & Wu, 2004).

Assessing Person Fit and Retest Effects at the Individual Level

The procedures proposed in the section above are based upon a number of assumptions, the fulfillment of which needs to be assessed if individual results are to be validly interpreted. This section discusses what are possibly the main assumptions as well as procedures for assessing whether they are met.

Assessing person fit

The procedures described in equations (16) to (22) assume that the $\mathbf{x}_i$ response vector behaves according to the 2WMIM solution that was fitted at the group level. It is indeed assumed that the fit at the group level was found to be appropriate. However, an acceptable overall model-data fit is still compatible with a certain proportion of individuals whose response patterns cannot be adequately explained by the model and, therefore, whose trait estimates and change estimates cannot be validly interpreted. Given this result, it is recommended that person fit must be routinely checked before individual results are interpreted (e.g. International Test Commission, 2014). This recommendation, however, is still far from being standard practice in IRT applications, and is virtually inexistent in SEM-based studies.

Ferrando (2014) proposed an IRT approach for assessing person fit based on test-retest data and discrete item scores. It can be readily adapted to the present scenario, and then further extended to the continuous case. Define first the log-likelihood index l_0-*rts* as the logarithm of the likelihood function in (16) evaluated using the $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ EAP point estimates instead of the unknown 'true' trait levels.

$$l_{0-rts}(\hat{\boldsymbol{\theta}}_i) = \sum_j^n \ln P(X_{ij}^{(1)} \mid \hat{\boldsymbol{\theta}}_i) + \sum_j^n \ln P(X_{ij}^{(2)} \mid \hat{\boldsymbol{\theta}}_i). \tag{23}$$

The index proposed by Ferrando (2014) for the binary and graded-response cases is a standardized likelihood-based index with the general form:

$$l_{z-rts} = \frac{l_{0-rts} - E(l_{0-rts})}{\sqrt{Var(l_0 - rts)}} \quad . \tag{24}$$

The conditional expectations and variances are given in the appendix. If the 'true' trait levels were known and used in equation (23), then the standardized index in (24) would be expected to asymptotically follow a standard normal distribution under the null hypothesis of consistency (Ferrando, 2014).

In the case of continuous responses, the log-likelihood index proposed by Ferrando (2007) can be extended and used with the 2WMIM. In this case, it takes the form:

$$l_{co-rts}(\hat{\boldsymbol{\theta}}_i) = \sum_j^n z_{ij}^{2(1)} + \sum_j^n z_{ij}^{2(2)} \tag{25}$$

where

$$z^{(m)}{}_{ij} = \frac{X^{(m)}{}_{ij} - \mu^{(m)}{}_j - \lambda^{(m)}{}_j \hat{\theta}_{im}}{\sigma^{(m)}{}_{\varepsilon j}} \tag{26}$$

By extending the results in Ferrando (2007) it follows that under the null hypothesis of consistency, the expected distribution of l_{co-rts} is χ^2 with 2(*n-1*) degrees of freedom. So, a normal approximation can be obtained as

$$l_{zco-rts}(\hat{\boldsymbol{\theta}}_i) = \sqrt{2l_{co-rts}(\hat{\boldsymbol{\theta}}_i)} - \sqrt{4n-5}. \qquad (27)$$

Indices of the type (24) and (27) are practical indices that assess the null hypothesis that the pattern is consistent against no specific alternative and are intended to be used as broad-screening tools for flagging potentially inconsistent patterns. As far as their interpretation and critical values are concerned, both (24) and (27) are interpreted (approximately) as standard normal *z* scores. In the case of (24) a large negative value is an indicator of inconsistency. As for *lzc*, it functions in the opposite direction. So, large positive values are indicators of misfit. For practical use, a standard cut-off value of -2 for index (24) and of +2 for (27) are expected to work well in principle for a test of reasonable length.

Assessing retest effects at the individual level.

In many T-R studies of the type considered here, REs are expected to exist and affect different individuals differently. At the overall level, addressing REs via correlated residuals is expected to lead to correct measurement and structural parameter estimates (Pitts, West & Tein, 1996). However, this modeling does not solve the potential problems that REs can cause in individual estimation.

Consider again the likelihood function (16). It is obtained by assuming that item responses are locally independent for fixed $\boldsymbol{\theta}_i$, and this local independence is assumed for: (a) responses to different items within a single measurement occasion; (b) repeated responses to the same items at Time-1 and Time-2; and (c) responses to different items on different occasions. If REs operate, part (b) of the assumption is violated, and this violation might lead to both incorrect trait estimates and incorrect standard errors. This issue is discussed in detail in Ferrando (2002, 2014) and will only be summarized here. Regarding the trait estimates, $\hat{\theta}_{i1}$ is expected to be correct. However, $\hat{\theta}_{i2}$ is expected to be biased towards $\hat{\theta}_{i1}$ and, therefore, the $|\hat{\delta}_i|$ change estimate in (19) biased towards zero (i.e. attenuated). As for the standard error, it is expected to be downwardly biased.

REs are expected to affect not only the individual estimates, but also the person fit measures (24) and (27), which are expected to be outwardly biased for the following reasons. If the individual responds consistently at Time-1 and tends to duplicate responses at Time-2, then both (24) and (27) would tend to flag this individual as more consistent than he/she really is. Conversely, inconsistent responding at Time-1 and retest effects would lead to the individual being flagged as more inconsistent than he/she really is.

The procedure proposed for assessing REs at the individual level is a refinement and extension of a previous proposal by Ferrando (2014), which, in turn is based on Yen's (1993) *Q3* rationale. Consider the following residual scores

$$s^{(1)}{}_{ij} = \frac{X^{(1)}{}_{ij} - E(X^{(1)}{}_{j} \mid \hat{\boldsymbol{\theta}}_i)}{\sqrt{Var(X^{(1)}{}_{j} \mid \hat{\boldsymbol{\theta}}_i)}} \quad ; \quad s^{(2)}{}_{ij} = \frac{X^{(2)}{}_{ij} - E(X^{(2)}{}_{j} \mid \hat{\boldsymbol{\theta}}_i)}{\sqrt{Var(X^{(2)}{}_{j} \mid \hat{\boldsymbol{\theta}}_i)}}, \tag{28}$$

which are based on the conditional expectations (13) to (15) but evaluated using the EAP estimates. Next, consider the n×1 vector $\mathbf{s}^{(1)}{}_{\mathrm{i}}$ containing the residual scores of respondent i for the n items at Time-1, and let $\mathbf{s}^{(2)}{}_{\mathrm{i}}$ be the corresponding vector at Time-2. The index *rtiQ3* is the product-moment correlation between $\mathbf{s}^{(1)}{}_{\mathrm{i}}$ and $\mathbf{s}^{(2)}{}_{\mathrm{i}}$: $rtiQ3_i = r\,(\mathbf{s}^{(1)}{}_{\mathrm{i}}, \mathbf{s}^{(2)}{}_{\mathrm{i}})$, and its rationale is as follows. If the model is correct and local independence for repeated responses holds, then the residuals in (28) are random error scores, and the expected value of *rtiQ3* for the individual is zero. On the other hand, if REs are operating, *rtiQ3* is expected to be positive and increase with the strength of the REs. From a practical point of view, a reasonable cutoff value for deciding whether REs impact the responses of the individual can be obtained by using the familiar Fisher's *z* transform

$$z_i(rtiQ3_i) = \frac{\frac{1}{2}\ln\frac{1+rtiQ3_i}{1-rtiQ3_i}}{1/\sqrt{n-3}} \quad , \tag{29}$$

and setting a one-tailed critical value for *z* of say +2.0. For a respondent who is detected to have been impacted by REs, the magnitude of the impact can then be ascertained by directly inspecting the *rtiQ3* value.

A Proposed Multi-Stage Approach

This section, which is intended for the more practically-oriented readers, aims to provide guidelines on how the procedures proposed so far are intended to be used in a test-retest study based on the 2WMIM.

- ***Stage 1: Fitting the 2WMIM***. A solution for the 2WMIM is specified (e.g. a strongly invariant solution with correlated residuals) and fitted in an appropriate sample. If the fit is considered to be acceptable, and the parameter estimates are accurate, then the measurement parameters (thresholds/intercepts, loadings and residuals variances) and the structural parameters (group means and variances and coefficient of stability) are taken as fixed and known, and used for scoring and assessing respondents in the subsequent stages. Group mean differences in change can also be assessed by using equations (7) and (8).

- ***Stage 2: obtaining individual scores and assessing individual appropriateness.*** EAP scores and the corresponding standard errors as proposed in equations (16) to (20) are obtained for each respondent. Next, these scores are used together with the fixed estimates in stage 1 for computing the person fit index (24) or (26) and the *rtiQ3* index in (28). If the person-fit results are acceptable and the REs are found to be negligible or weak, then the individual assessment procedures can be applied to this respondent. If this is not the case, further post-hoc assessments can be made to ascertain the causes of misfit.

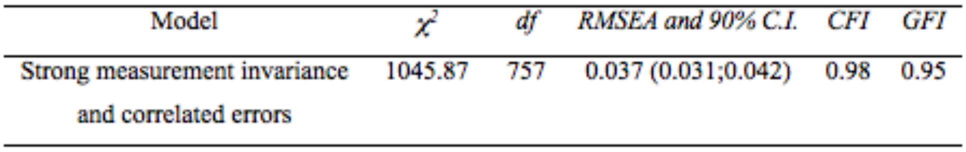
- ***Stage 3: individual estimation of change***. For those individuals (hopefully the majority of the group) who respond consistently with the model and are not greatly affected by REs, valid and accurate estimates of change can be obtained by using the results in equations (21) (significance and accuracy of the change estimate) and (22) (individual effect size).

ILLUSTRATIVE EXAMPLE

The functioning of the present proposal is illustrated with a dataset from the experimental-faking study by Ferrando and Anguiano-Carrasco (2009). A group of 277 undergraduates was administered a Lie scale consisting of 20 binary items at two points of time with a retest interval of 6 weeks. At Time-1 the participants were asked to respond under standard instructions. At Time-2 they were asked to imagine themselves as job applicants and try to give a good impression when answering regardless of the truthful answer. So, the design was: pretest-treatment-posttest. The Lie scale chosen was considered to be a measure of the impression management (IM) construct (Paulhus 1991), and the expected result was a strong mean shift at Time-2 toward more socially desirable responding (i.e. higher mean trait levels of IM at Time-2).

Given the binary format of the items, the 2WMIM that was fitted used the link function in equation (1), which means that each measurement submodel in equation (3) was an IRT two-parameter normal-ogive model. The solution that was specified was strongly invariant with correlated over-time residuals, and was fitted using WLSM estimation as implemented in the Mplus program version 5.1 (Muthén, & Muthén, 2007). A reduced example of the program code can be found in the appendix. The goodness-of-fit results are shown in table 1 and indicate that the overall model-data fit is quite acceptable.

Table 1. Goodness of fit Results for the 2WMIM.

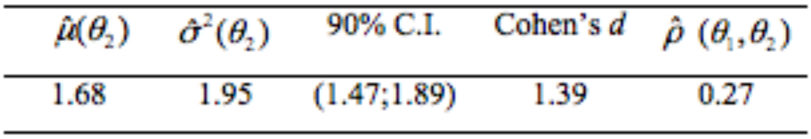


Model	χ^2	*df*	*RMSEA and 90% C.I.*	*CFI*	*GFI*
Strong measurement invariance and correlated errors	1045.87	757	0.037 (0.031;0.042)	0.98	0.95

Note. χ^2: WLSM chi-square goodness-of-fit statistic; d.f.: degrees of freedom; RMSEA and 90% C.I.: point and interval estimate of the root mean squared error of approximation; CFI: comparative fit index; GFI: goodness of fit index.

Results on the measurement submodels in equation (3) can be summarized as follows: (a) the thresholds ranged from -0.6 to 1.5 with a mean of 0.18 (i.e. a good spread of item locations), and (b) the standardized loadings ranged from 0.3 to 0.7 with a mean of 0.52 (i.e. moderately discriminating items). Results on the structural submodel in (4) are in table 2, and can be summarized in two points. First, at the group level there is a clear change in the expected direction. This change is significant, can be assessed quite accurately (i.e. a relatively narrow confidence interval), and the effect size can be qualified as strong (Cohen, 1988). Second, the coefficient of stability is low, and the variance of the difference scores is high, which means that there are substantial individual differences in the magnitude of change. Given these results, it is clearly necessary to study changes at the individual level.

Table 2. Parameter estimates of the structural submodel.



$\hat{\mu}(\theta_2)$	$\hat{\sigma}^2(\theta_2)$	90% C.I.	Cohen's *d*	$\hat{\rho}\ (\theta_1,\theta_2)$
1.68	1.95	(1.47;1.89)	1.39	0.27

Note. $\hat{\mu}(\theta_2)$: mean group trait level estimate at Time-2; $\hat{\sigma}^2(\theta_2)$: variance estimate at Time-2; 90% C.I.: confidence interval for the mean estimate; $\hat{\rho}\ (\theta_1,\theta_2)$: coefficient of stability.

MATLAB (1999) routines were written to compute the procedures proposed in this article, and some of them are provided in the appendix. The general results will first be summarized, and then the approach will be illustrated by using the results from three participants. As for the summary, the group mean of $l_{z\text{-}rts}$ in (24) was -0.25, which is compatible with a small proportion of inconsistent respondents in this group: if the -2 cut-off value proposed above is used, then 30 (10.8%) respondents would be flagged as potentially inconsistent. As for REs, the mean of *rtiQ3* was 0.08, thus suggesting that for most respondents they have little impact. Again, if the +2 cut-off value proposed here is used, only 24 (8.7%) respondents would be considered to be impacted by REs. Overall, the results suggest that estimates of change are valid and meaningful for most of the respondents in the group, but non-informative or non-valid for some (see below).

The EAP estimated trait levels of respondent n° 92 were $\hat{\theta}_{92,1} = 0.10$ and $\hat{\theta}_{92,2} = 0.83$. On the basis of these estimates, the $l_{z\text{-}rts}$ index was computed, and gave a value of $l_{z\text{-}rts}$(92)= -4.10, which indicates a substantial degree of person misfit for this response pattern. To obtain more information, the basic *lz* index was next computed separately in the Time-1 and the Time-2 sub-patterns, and the values obtained were -2.63 (Time-1) and -3.14 (Time-2) which indicate inconsistent responding on both occasions. In addition, the value of the *rtiQ3* index was *rtiQ3*(92)=0.03, clearly nonsignificant, thus suggesting that the impact of REs was negligible for this respondent. So, REs cause no outward bias in $l_{z\text{-}rts}$ in this case. Given the clear inconsistency of his responses, the trait estimates of respondent n° 92 cannot be validly interpreted, and it seems meaningless to continue the analysis in order to assess change.

The second illustration corresponds to respondent n° 213, whose trait estimates were $\hat{\theta}_{213,1} = 0.32$ and $\hat{\theta}_{213,2} = 0.38$. The value of $l_{z\text{-}rts}$ in this case was $l_{z\text{-}rts}$(213)= 0.92, which indicates responding which is quite consistent with the model. The value of *rtiQ3*, however, was *rtiQ3*(213)=0.82, very significant (*zrtiQ3*(213)=4.77) and very high, which indicates that the responses of this participant were strongly impacted by REs. The most likely interpretation for these results can be summarized in two points. First, the respondent is consistent, but the strong REs cause an outward bias in $l_{z\text{-}rts}$ that makes her appear more consistent than she really is. Second, the trait estimate at Time-1 is probably correct and can be validly interpreted. However, the trait estimate at Time-2 is probably biased towards $\hat{\theta}_{213,1}$ so the small change estimate for this respondent ($\hat{\delta}_{213} = 0.06$) is probably attenuated and cannot be validly interpreted.

The final illustration corresponds to respondent nº 5, with trait estimates $\hat{\theta}_{5,1} = -0.63$ and $\hat{\theta}_{5,2} = 2.18$. The appropriateness values were $l_{z\text{-}rts}(5)= 0.01$, and *rtiQ3*(5)=0.01, both clearly nonsignificant. So, the responses of nº 5 can be considered to be appropriate and free from REs, and so, they can be validly interpreted.

The individual amount of change estimate for this respondent was $\hat{\delta}_5 = 2.81$. The corresponding 90% confidence (credibility) interval obtained according to (21) was: (1.74; 3.88), and the individual effect size measure in (22) was 2.31. The interpretation of these results is that there is a strong and statistically significant change between the Time-1 and the Time-2 trait levels that goes in the expected direction. As expected, this change cannot be measured as accurately as the group-level mean change in table 2 (the confidence interval is far wider here). Even so, the change estimated for this individual is substantially larger than the average change estimated for the entire group (see table 2).

DISCUSSION

The main purpose of this article is to extend a simple and well known SEM in order to obtain additional information that might be useful when individual assessment is of interest in a T-R study. So, the present proposal is a potentially relevant contribution at the substantive level. Furthermore, as one reviewer pointed out, the methods proposed here could also be used in developmental studies based on far longer retest intervals.

The procedures proposed in this article require a moderately large set of items to work properly, and this is possibly their main limitation as far as applicability is concerned. On the one hand, individual estimation of change is very imprecise in short tests (Kruyen, Emons & Sijtsma, 2013). On the other, the longer the response pattern is, the better person-fit measures and individual retest indices work (Ferrando, 2014). These principles are generally clear in IRT, but are not so clear in SEM. Possibly because of technical limitations, most 2WMIM applications of the type considered here that were made during the 1980s and 1990s were based on small sets of 3 to 8 items (e.g. Aish & Jöreskog, 1990, Muthén, 1984). This makes the present proposal unfeasible in practice. In recent decades, however, the procedures for estimating and testing SEMs have made considerable progress, specially in the analysis of categorical variables (e.g. Muthén & Muthén, 2007), and it is now possible to fit solutions based on a moderate-to-large number of items (say between 20 and 60). So, solutions based on reasonably long tests are expected to be more and more common in the future, and, if they are, the methods proposed here are expected to become more and more relevant. The appropriateness of the present procedure needs to be assessed with tests of different lengths and further intensive research, mostly based on simulation, needs to be carried out. More generally, simulation studies will be needed to assess how the procedures work under different conditions. This is a clear aim for further studies.

The present proposal can be further extended in other directions and many aspects improved. To start with, the proposal has only considered the unidimensional case. Technically, extending all the procedures proposed here to the multidimensional case does not present any particular problems. However, the resulting models then become more complex, and are potentially more unstable. Overall, whether it is practically feasible to apply the present proposal to multidimensional questionnaires requires further research based on both simulation and real data.

A second clear extension is multiple-group analysis because, for example, most pretest-posttest studies also use a control group (e.g. Ferrando & Anguiano-Carrasco, 2009). As occurs with multidimensional extensions, multiple-group extensions do not pose special problems other than that they increased model complexity, so they must be considered in the future.

I shall now discuss two of the many points that can be improved and developed further. First, the present proposal can be viewed as three-stage. Mean/threshold vectors and correlation/covariance matrices are obtained in the first stage. Measurement and structural estimates from the SEM are then obtained in the second stage. And, finally, the SEM estimates are taken as fixed and known so that individuals can be assessed in the third stage. So, overall, uncertainty in the estimates in each stage is ignored and can propagate to subsequent stages. Thus, the second-stage SEM estimates cannot be stable and accurate if the correlations which serve as the input (first stage) are not (Lorenzo-Seva & Ferrando, 2015). And the uncertainty that has accumulated is ignored when scoring individuals and assessing score appropriateness. In strong, clearly structured solutions fitted in large samples this point has probably little practical relevance. In other scenarios, however, procedures that take into account parameter uncertainty might be considered (e.g. Yang, Hansen & Cai, 2012).

With regards to the second point, the procedures proposed here to determine cut-off values in person-fit and RE assessments rely on approximations to known distributions. For tests of reasonable length these approximations are expected to suffice in practice. However, they can be clearly improved (see Ferrando, 2014).

In spite of the limitations acknowledged above, the results obtained in the illustrative study are clearly encouraging. The procedures worked well, and made it possible to detect problematical patterns that could have led to misleading interpretations. Conversely, the fact that most of the patterns behaved in accordance with the normative model provided support for a valid interpretation of the estimates. Finally, for the reasons discussed above, the individual estimates of trait levels and change can be regarded as more accurate and correct than the standard procedures used in this type of design.

Some of the procedures proposed in this article cannot be carried out by using standard SEM programs and require specific software to be developed. More in detail, if the present proposal is to be put to widespread use, a free user-friendly program for (a) obtaining individual estimates of change, (b) computing person fit indices, and (c) assessing REs at the individual level must be available. This is also a clear aim for future developments.

RESUMEN

Un enfoque combinado TRI-MEE para la evaluación individual en estudios test-retest. El modelo longitudinal de medida con múltiples indicadores evaluados en dos ocasiones (MI2O) se utiliza habitualmente en estudios test-retest y proporciona información al nivel de grupo y al nivel de ítems. Además, cuando se aplica a respuestas binarias o graduadas dicho modelo se convierte en una extensión de algunos modelos básicos de teoría de respuesta al ítem (TRI). En este artículo se explotan las relaciones TRI-MI2O para obtener información adicional al nivel de cada individuo. Se proponen procedimientos para (a) obtener estimaciones individuales de niveles en el rasgo y magnitud del cambio, y (b) evaluar si se cumplen o no los supuesto básicos en que dichas estimaciones se fundamentan. Los procedimientos propuestos se organizan en un marco general que puede utilizarse con respuestas binarias, graduadas o continuas. Se discute la relevancia de la propuesta y se proponen recomendaciones para utilizarla en investigación aplicada. Finalmente, la propuesta se ilustra con un ejemplo empírico donde funcionó bien y proporcionó información útil.

REFERENCES

Aish, A. M., & Jöreskog, K. G. (1990). A panel model for political efficacy and responsiveness: An application of LISREL 7 with weighted least squares. *Quality and Quantity*, 24, 405-426.

Andrade, DF, & Tavares, HR (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *95*, 1–22. doi: 10.1016/j.jmva.2004.07.005

Bock, R.D. & Mislevy, R.J. (1982). Adaptative EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6,* 431-444.

Bollen, K.A. & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In P.V. Marsden (ed.) *Sociological Methodology 1991* (pp. 235-262). New York: Basil Blackwell.

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.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model . *Psychometrika, 58*, 37-52. doi: 10.1007/BF02294469

Cohen, J. (1988) *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Embretson, SE (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515. doi: 10.1007/BF02294487

Ferrando, P.J. (2002). An IRT-based two-wave model for studying short-term stability in personality measurement. *Applied Psychological Measurement*, *26*, 286-301. doi: 10.1177/0146621602026003004

Ferrando, P.J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research*, *42*, 481-508. doi: 10.1080/00273170701382583

Ferrando, P.J. (2014). A Comprehensive approach for assessing Person Fit with Test–Retest data. *Educational and Psychological Measurement, 74*, 585-610. doi:10.1177/001316441351855.

Ferrando, P.J. (2015). Assessing retest effects at the individual level: A general IRT-based approach. *Psicológica, 36*, 141-161.

Ferrando, P.J., & Anguiano-Carrasco, C. (2009). Assessing the Impact of Faking on Binary Personality Measures: An IRT-Based Multiple-Group Factor Analytic Procedure. *Multivariate Behavioral Research, 44*, 497-524. doi:10.1080/00273170903103340

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. doi: 10.1177/0146621609344844

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3-26. doi: 10.1177/0146621602239474

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling*, *20*, 504-517. doi: 10.1080/10705511.2013.797837

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. doi: 10.1037/0021-9010.87.2.243

International Test Commission (2011). *ITC guidelines for quality control in scoring, test analysis, and reporting of test scores*. http://intestcom.org.

Jöreskog, K.G. (1979). Statistical Estimation of Structural Models in Longitudinal-Developmental Investigations. In J.R. Nesselroade, & P.B. Baltes. (Eds.) *Longitudinal Research in the Study of Behavior and development*. Academic Press: New York.

Kenny, D.A., & Campbell, D.T. (1989). On the measurement of Stability in Over-Time Data. *Journal of Personality*, *57*, 445-481. doi: 10.1111/j.1467-6494.1989.tb00489.x

Kruyen, P.M., Emons, W.H.M., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement*, *38*, 201-216. doi: 10.1177/0146621613510061

Little, T.D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford Press.

Little, T.D., Slegers, D.W., & Card, N.A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72. doi: 10.1207/s15328007sem1301_3

Lorenzo-Seva, U. & Ferrando, P.J. (2015). POLYMAT-C: a comprehensive SPSS program for computing the polychoric correlation matrix. *Behavior Research Methods, Instruments & Computers* (in press; available at: http://link.springer.com/article/10.3758/s13428-014-0511-x).

MATLAB. (1999). *MATLAB 5.3 Release 11.1*. Natick MA: The Math Works Inc.

Mellenbergh, G.J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, *23*, 87 – 89 doi: 10.1177/01466219922031211

Millsap, R.E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R.C. MacCallum (Eds.) *Factor analysis at 100* (pp. 131-152). Mahwah: LEA.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115-132. doi: 10.1007/BF02294210

Muthén, L.K., & Muthén, B. (2007). *Mplus user's guide*. *Fifth Edition*. Los Angeles: Muthén & Muthén.

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver & L.S. Wrightsman (eds.) *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego: Academic Press.

Pitts, S.C., West, S.G., & Tein, J. (1996). Longitudinal measurement models in evaluation research: examining stability and change. *Evaluation and Program Planning, 19*, 333-350. doi: 10.1016/S0149-7189(96)00027-4

Raykov, T. and Penev, S. (2002). Exploring structural equation model misspecifications via latent individual residuals. In G.A. Marcoulides and I. Moustaki (eds.) *Latent variable and latent structure models* (pp. 121-134). Mahwah, NJ: LEA.

Reise, S.P. & Widaman, K.F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods*, *4*, 3-21. doi: 10.1037/1082-989X.4.1.3

Reise, S.P., & Haviland, M.G. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84,* 228-238. doi : 10.1207/s15327752jpa8403_02

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*. doi:10.1007/BF02290599

Wang, W.-C., & Wu, C.-I. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement*, *64*, 758-780. doi:10.1177/0013164404264118

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375. doi:10.1111/j.1745-3984.1984.tb01040.x

Weiss, D.J. & Von Minden, S. (2011). Measuring Individual Growth with Conventional and Adaptive Tests. *Journal of Methods and Measurement in the Social Sciences, 2,* 80-101.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*, 264-290. doi: 10.1177/0013164411410056

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. doi: 10.1177/014662168400800201

APPENDIX

Expected Values and Variances for the Log-Likelihood Person Fit Indices

For the three types of item responses considered in the article, the likelihood function (16) evaluated with the $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ EAP point estimates is.

Binary:

$$L(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}_i) = \prod_{j}^{n} P(X^{(1)}_{ij} = 1 \mid \hat{\theta}_{i1})^{X^{(1)}_{ij}} (1 - P(X^{(1)}_{ij} = 1 \mid \hat{\theta}_{i1}))^{1-X^{(1)}_{ij}} \cdot \prod_{j}^{n} P(X^{(2)}_{ij} = 1 \mid \hat{\theta}_{i2})^{X^{(2)}_{ij}} (1 - P(X^{(2)}_{ij} = 1 \mid \hat{\theta}_{i2}))^{1-X^{(2)}_{ij}} \tag{30}$$

Graded:

$$L(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}_i) = \prod_{j}\prod_{r} P(X^{(1)}_{ij} = r \mid \hat{\theta}_{i1})^{u^{(1)}_{ijr}} \prod_{j}\prod_{r} P(X^{(2)}_{ij} = r \mid \hat{\theta}_{i2})^{u^{(2)}_{ijr}} \tag{31}$$

where u_{ijr}=1 if respondent *i* chooses category *r* for item *j*, and u_{ijr}=0 otherwise.

Continuous:

$$L(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}_i) = \prod_j \left[\frac{1}{\sigma^{(1)}{}_{\varepsilon j}\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{X^{(1)}{}_{ij} - \mu^{(1)}{}_j - \lambda^{(1)}{}_j \hat{\theta}_{i1}}{\sigma^{(1)}{}_{\varepsilon j}} \right)^2 \right] \cdot \prod_j \left[\frac{1}{\sigma^{(2)}{}_{\varepsilon j}\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{X^{(2)}{}_{ij} - \mu^{(2)}{}_j - \lambda^{(2)}{}_j \hat{\theta}_{i2}}{\sigma^{(2)}{}_{\varepsilon j}} \right)^2 \right] \qquad (32)$$

Next, by using the results in equations (10) to (15), the conditional means and variances of the likelihood-based person fit indices for the binary and the graded response cases are found to be

Binary:

$$\begin{aligned} &E(l_{0-rts}(\hat{\boldsymbol{\theta}}_i)) = \\ &\sum_{j=1}^{n} \{P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}) \ln P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}) + \Big[(1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}) \ln(1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}))\Big]\} \\ &+ \sum_{j=1}^{n} \{P(X^{(2)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i2}) \ln P(X^{(2)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i2}) + \Big[(1 - P(X^{(2)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i2}) \ln(1 - P(X^{(2)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i2}))\Big]\} \end{aligned} \qquad (33)$$

and

$$\begin{aligned} Var(l_{0-rts}(\hat{\boldsymbol{\theta}}_i)) = &\sum_{j=1}^{n} \{P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})(1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}) \left[\ln \frac{P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})}{1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})} \right]^2 \} \\ &+ \sum_{j=1}^{n} \{P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})(1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1}) \left[\ln \frac{P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})}{1 - P(X^{(1)}{}_{ij} = \mathit{1} \mid \hat{\theta}_{i1})} \right]^2 \} \end{aligned} \qquad (34)$$

Graded:

$$\begin{aligned} E(l_{0-rts}(\hat{\boldsymbol{\theta}}_i)) = &\sum_j \sum_r \{P(X^{(1)}{}_{ij} = r \mid \hat{\theta}_{i1}) \ln P(X^{(1)}{}_{ij} = r \mid \hat{\theta}_{i1})\} \\ &+ \sum_j \sum_r \{P(X^{(2)}{}_{ij} = r \mid \hat{\theta}_{i2}) \ln P(X^{(2)}{}_{ij} = r \mid \hat{\theta}_{i2})\} \end{aligned} , \qquad (35)$$

and

$$
\begin{aligned}
&Var(l_{0-rts}(\hat{\boldsymbol{\theta}}_i)) = \\
&\sum_j\sum_r\left[\sum_s P(X^{(1)}{}_{ij}=r\,|\,\hat{\theta}_{i1})P(X^{(1)}{}_{ij}=s\,|\,\hat{\theta}_{i1})\ln P(X^{(1)}{}_{ij}=r\,|\,\hat{\theta}_{i1})\ln(\frac{P(X^{(1)}{}_{ij}=r\,|\,\hat{\theta}_{i1})}{P(X^{(1)}{}_{ij}=s\,|\,\hat{\theta}_{i1})})\right] \\
&+\sum_j\sum_r\left[\sum_s P(X^{(2)}{}_{ij}=r\,|\,\hat{\theta}_{i2})P(X^{(2)}{}_{ij}=s\,|\,\hat{\theta}_{i2})\ln P(X^{(2)}{}_{ij}=r\,|\,\hat{\theta}_{i2})\ln(\frac{P(X^{(2)}{}_{ij}=r\,|\,\hat{\theta}_{i2})}{P(X^{(2)}{}_{ij}=s\,|\,\hat{\theta}_{i2})})\right].
\end{aligned}
\tag{36}
$$

Example of Mplus code for fitting the T-R model with strong invariance and correlated residuals (with output details).

```
TITLE:  two-wave model for dichotomous responses. Example with 5 items
DATA: FILE IS c:lietr5.dat;
    FORMAT IS  10F2.0;
VARIABLE: NAMES ARE t1-t5 r1-r5;
      CATEGORICAL = t1-t5 r1-r5;
ANALYSIS:  TYPE = MEANSTRUCTURE;
         ESTIMATOR = WLSM;
MODEL:
      f1 BY t1* t2-t5;
      f2 BY r1* r2-r5;
      f2 ON f1;
      f1@1;
      [f1@0];
      [f2*];
      f1 BY t1(1);
      f2 BY r1(1);
      f1 BY t2(2);
      f2 BY r2(2);
      f1 BY t3(3);
      f2 BY r3(3);
      f1 BY t4(4);
      f2 BY r4(4);
      f1 BY t5(5);
      f2 BY r5(5);
      [t1$1](6);
      [r1$1](6);
      [t2$1](7);
      [r2$1](7);
      [t3$1](8);
      [r3$1](8);
      [t4$1](9);
      [r4$1](9);
      [t5$1](10);
      [r5$1](10);
      t1-t5 PWITH r1-r5;
OUTPUT: standardized residual;
```

Output details (latent variable sub-model)

```
Means
   F1              0.000      0.000     (mean at Time-1 fixed to zero)

 Intercepts
   F2              1.239      0.151      8.195      0.000 (intercept equals mean at Time-2, see
                                                              (equation 5).
 Variances
   F1              1.000      0.000     (variance at Time-1 fixed to one).
 Residual Variances
   F2             1.692     0.312      5.416      0.000 (residual variance at Time-2, equation 5).
F2       ON
   F1            0.389     0.149      2.604      0.009 (slope of the latent-variable sub-model, see
equation 4).
```

Example of Matlab code for computing the global person fit index

```
% lzret computes the standardized likelihood statistic for retest
% data and binary responses when strong invariance is assumed.
%
function [l, lz] = lzret(filat,filart,tt,tr, LAM,thres, sige1,sige2)

% INPUT
%
%   filat ->  row vector of responses at Time-1
%   filart -> row vector of responses at Time-2
%   tt ->  trait estimate at Time-1
%   tr -> trait estimate at Time-2
%   LAM -> pattern matrix (dimension items x factors).
%  thres -> vector of item thresholds
%  sige1 -> residual standard deviations at Time-1
%  sige2 -> residual standard deviations at Time-2
%
% OUTPUT
%
% l-> raw likelihood index
% lz -> standardized likelihood index
[n,m]= size (LAM);
  l=0.0;
  med=0.0;
  var=0.0;
  pt=0.0;
  prt=0.0;
  for i=1:n,
% Computes the conditional probabilities

     pt = pconmul(tt,LAM(i,:),thres(i),sige1(i));
     prt = pconmul(tr,LAM(i,:),thres(i),sige2(i));

% Computes the log-likelihood and its expectations

     temp1=filat(i).*log(pt)+(1-filat(i)).*log(1-pt);
     temp2=filart(i).*log(prt)+(1-filart(i)).*log(1-prt);
     l = l+temp1+temp2;
     temp3= (pt.*log(pt))+((1-pt).*log(1-pt));
     temp4= (prt.*log(prt))+((1-prt).*log(1-prt));
     med= med+temp3+temp4;
     temp5= pt.*(1-pt).*log(pt/(1-pt)).*log(pt/(1-pt));
     temp6= prt.*(1-prt).*log(prt/(1-prt)).*log(prt/(1-prt));
     var=var+temp5+temp6;
     pt=0.0;
     prt=0.0;
  end;

% Computes the standardized index

  sd=sqrt(var);
  lz = (l-med)/sd;
return;
```

Example of Matlab code for computing the rtiQ3 index

```
function [rtiQ3, zQ3] = q3bin(filat,filart, tt, tr, LAM,thres,sige1,sige2)
% q3bin computes the rtiQ3 index and the corresponding standardized value
%
% INPUT
%   filat ->  row vector of responses at Time-1
%   filart -> row vector of responses at Time-2
%   tt ->  trait estimate at Time-1
%   tr -> trait estimate at Time-2
%   LAM -> pattern matrix (dimension items x factors).
%  thres -> vector of item thresholds
%  sige1 -> residual standard deviations at Time-1
%  sige2 -> residual standard deviations at Time-2
%
% OUTPUT
% rtiQ3 -> index
% zQ3 -> standardized value

[n,m]= size (LAM);
```

```
  col1=zeros(n,1);
  col2=zeros(n,1);
  vt=zeros(n,1);
  vrt=zeros(n,1);
  siget=zeros(n,1);
  sigert=zeros(n,1);

for i=1:n,

% Computes the standardized residual scores

      pt(i) = pconmul(tt,LAM(i,:),thres(i),sige1(i));
      prt(i) = pconmul(tr,LAM(i,:),thres(i),sige2(i));
      vt(i)= pt(i)* (1-pt(i));
      vrt(i)= prt(i)* (1-prt(i));
      siget(i)=sqrt(vt(i));
      sigert(i)=sqrt(vrt(i));
      col1(i)=(filat(i)-pt(i))/siget(i);
      col2(i)=(filart(i)-prt(i))/sigert(i);
  end;

% computes the correlation between the residuals (rtiQ3) and its standardized
% value based on Fisher's z transform.

  totr= corrcoef(col1,col2);
  tmp4= totr(1,2);
  zf= 0.5*log((1+tmp4)/(1-tmp4));
  deno=1/sqrt(n-3);
  tmp5=zf/deno;
return;
```

(Manuscript received: 8 January 2015; accepted: 1 June 2015)