Psicológica (2010), 31, 319-334.

Assessing short-term individual consistency using IRT-based statistics

Pere J. Ferrando*

Rovira i Virgili University, Spain

This article proposes a procedure, based on a global statistic, for assessing intra-individual consistency in a test-retest design with a short-term retest interval. The procedure is developed within the framework of parametric item response theory, and the statistic is a likelihood-based measure that can be considered as an extension of the well-known *lz* person-fit index. The rationale for using and interpreting the proposed statistic is discussed, and an adapted standardized residual at the item level is also proposed to obtain clues about the possible causes of the detected inconsistency. The procedure is illustrated with a real-data example and a parallel simulation in the personality domain.

In recent decades in applied psychometrics interest in the assessment of the intra-individual consistency of the responses over a set of test items has been growing. Within the framework of parametric item response theory (IRT), which is the one considered here, most of the research on this topic has focused on developing and evaluating global statistics (i.e. scalar measures based on the complete pattern of item responses) that measure the extent to which the answering behavior of a respondent is consistent with the psychometric model which was fitted to the data. These statistics were initially known as appropriateness measures (Levine & Rubin, 1979), but now they are usually referred to as person-fit indices (e.g. Meijer & Sijtsma, 1995, 2001). Most of these model-based statistics can be classified in two broad classes (see Meijer & Sijtsma, 1995). The measures in the first class are based on a residual that reflects the differences between the observed item scores and the scores expected from the model. The measures in the second

* This research was partially supported by a grant from the Spanish Ministry of Education and Science (PSI2008-00236/PSIC). Correspondence: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Departamento de Psicología. Carretera Valls s/n. 43007 Tarragona (Spain). E-mail: perejoan.ferrando@urv.cat

class are based on the likelihood function. In this second class, two of the best-known and most commonly used person-fit indices are the *lo* and *lz* measures initially proposed by Levine and Rubin (1979).

Person-fit indices were initially intended for the ability and aptitude domain and mainly for practical purposes. However, over time they have also been used in the personality domain (Ferrando & Chico, 2001, Reise, 1995, Reise & Flannery, 1996, Reise & Waller, 1993, Meijer et al., 2008). The assessment of intra-individual consistency in personality is more than a practical issue, because in this domain consistency is a topic of central theoretical relevance (e.g. Tellegen, 1988).

In personality measurement three types of intra-individual consistency are generally distinguished depending on the temporal framework in which the assessment takes place (e.g. Fiske & Rice, 1955, Lumsden 1977, Watson, 2004). The first type is momentary consistency, which is assessed by analyzing the responses of the individual over the set of items during a single test administration. The second type is short-term consistency, usually assessed by using a two-wave or a multi-wave retest design with a short retest interval (generally from a few days to a few weeks). Finally, the third type is long-term consistency, which is concerned with trait stability, and which is generally assessed by using two-wave or multi-wave designs with long retest periods (e.g. Conley, 1984). Cattell (1986) made a more specific differentiation and distinguished between “dependability” (a retest interval shorter than two months) and “stability” (a retest interval of two months or more). Now, according to these distinctions, standard person-fit indices appear to be useful measures for assessing the first type of consistency at the individual level (Reise, 1995, Reise & Flannery, 1996, Reise & Waller, 1993). However, statistics for measuring the other two types also seem to be of potential interest. In particular, this paper is concerned with the assessment of short-term individual consistency. In personality measurement, this assessment is of both theoretical and practical interest. Theoretically, the degree of short-term consistency can be considered partly as a property of the trait (Cattell, 1986), and, for a personality theorist the differences of traits in their degree of consistency is important information. In experimental and clinical settings, short-term designs are commonly used to assess the effects of experimental conditions or treatments. Also, in the selection domain, short-term designs are used to gauge the effects of test-coaching and practice.

The classical approach for assessing short-term consistency in personality is to assume that the individuals are perfectly stable over time: hence, all inconsistency is due solely to measurement error (see e.g.

Watson, 2004). However, this rationale is surely too simple. First, even with short retest intervals, individual transient fluctuations such as mood changes, cognitive energy variations, mental attitude changes, etc are expected to occur (Lumsden 1977, Schmidt & Hunter, 1996). Second, changes in the assessment conditions (e.g. instructions, pressures or motivating conditions) might give raise to 'true' or simulated temporary changes in the trait levels (e.g. Schuldberg, 1990, Zickar & Drasgow, 1996). Third, even if conditions remain the same over different administrations, the mere fact of having responded at Time 1 (prior exposure) might lead to systematic changes at the retest (e.g. Knowles, 1988). Finally, we also need to consider the retest effects, understood here as the tendency for individuals to duplicate their former responses (Gulliksen, 1950). This tendency might be due to memory effects, or to incidental item features which are unrelated to the trait, but which tend to elicit the same response on each occasion (Thorndike, 1951). So, overall, it seems more realistic to assume that short-term inconsistency arises as a result of a complex item × respondent interaction process (e.g. Ferrando, Lorenzo-Seva & Molina, 2001, Schuldberg, 1990). In spite of this assumption, however, most of the research on the causes of short-term inconsistent responding has focused solely on item characteristics (see Ferrando et al. 2001 for a review).

The present paper develops and proposes an IRT-based statistic for assessing short-term consistency at the individual level. This statistic is likelihood-based, and can be considered as an extension of the *lz* person-fit index mentioned above. It is intended for binary items which are calibrated with a unidimensional IRT model. Of the existing parametric IRT models, binary unidimensional models are the simplest and best known. So, they seem to be an appropriate starting framework for developing new measures. The number of questionnaires based on binary items is still substantial in personality, so the potential interest of the statistic in applied research seems clear.

The statistic proposed here is expected to be of both theoretical and practical interest. At the practical level, it can be used for flagging individuals for whom the estimated trait level at Time 1 might be inappropriate for making valid inferences in the short term. A second practical use is for detecting outliers (i.e. inconsistent individuals) in longitudinal studies. It is well known that the estimated parameters and fit results in structural equation models are easily distorted if outliers are present (e.g. Bollen & Arminger, 1991). At a more theoretical level, the statistic can provide additional information about the inconsistency of the response behavior of the individual beyond that provided in a single test administration, and can be useful for assessing different types of individual change as well as retest effects.

Rationale and Description of the *lz-ch* Statistic

Consider a personality test that measures a single trait θ and which is made up of $1,\ldots j,\ldots n$ items scored as 0 or 1. The item responses are assumed to behave according to a specific parametric IRT model, and are independent for fixed trait level (local independence). Let $P_j(\theta)$ be the item response function corresponding to the IRT model, and let $\mathbf{x}_i=(x_{i1}\ldots\ldots x_{in})$ be the response pattern of individual i. Then, the log likelihood of $\mathbf{x}_i$ is

$$L(\theta_i) = \sum_j^n x_{ij} \ln(P_j(\theta_i)) + (1 - x_{ij}) \ln(1 - P_j(\theta_i)) \qquad (1)$$

Furthermore, the mean and variance of $L(\theta_i)$ are (Drasgow, Levine & Williams, 1985)

$$E(L(\theta_i)) = \sum_j^n P_j(\theta_i) \ln(P_j(\theta_i)) + (1 - P_j(\theta_i)) \ln(1 - P_j(\theta_i)) \qquad (2)$$

and

$$Var\,(L(\theta_i)) = \sum_j^n P_j(\theta_i)(1 - P_j(\theta_i)) \left[\ln \frac{P_j(\theta_i)}{1 - P_j(\theta_i)} \right]^2 \qquad (3)$$

Levine and Rubin (1979) defined the *lo* person-fit index as the log likelihood (1) computed using the maximum likelihood (ML) estimate of θ_i. The rationale for this choice is that when a pattern is inconsistent, no value of θ makes the likelihood of this pattern large. So, even when evaluated at the maximum, the value of (1) is still relatively small.

A limitation of *lo* that can be noted by inspection of (2) is that its value depends generally on the trait level. Furthermore, the index has no immediate theoretical reference distribution for assessing the values obtained. To overcome these problems, Drasgow et al. (1985) proposed a standardized version of *lo*, which they named *lz*. It is given by

$$l_z = \frac{l_o - E(l_o)}{\left[Var\,(l_o)\right]^{1/2}} \tag{4}$$

where the mean and variance are obtained from (2) and (3) computed using the ML estimate of θ_i. The standardization provides a known scale and is intended to reduce the dependency of the values on θ. Furthermore, because the terms in (1) are independent for fixed θ, as the number of items increases the distribution of *lz* is expected to approach the standard normal. This approximation requires some conditions to be fulfilled (see below). As for interpretation, typically only the left tail of the distribution is considered: large negative values indicate potential person misfit. Large positive values would indicate an over-consistent, deterministic (i.e. Guttman-type) responding, however they are seldom interpreted.

Consider now that the test is administered to the same respondents at two points of time with a given retest interval. The present development is based on the additional assumption that the local independence principle also holds for repeated measurements of the same item, or, more specifically, that for fixed θ_i the conditional distributions of the responses to the same item in two repeated administrations are independent of each other. If this is so, the conditional probability of a response change to item *j* is given by

$$Pch_j(\theta_i) = 2P_j(\theta_i)(1 - P_j(\theta_i)) \tag{5}$$

i.e. the probability of the 1-0 pattern of change plus the probability of the 0-1 pattern (Ferrando et al., 2001, Nowakowska, 1983). We shall now define the indicator variable y_j to denote response change to item *j* as

$$y_{ij} = \begin{Bmatrix} 1,\ if\ response\ of\ item\ j\ was\ changed\ at\ Time2 \\ 0,\ otherwise \end{Bmatrix} \tag{6}$$

and $\mathbf{y}_i=(y_{i1}.....y_{in})$ as the “change” response pattern of individual *i.* Under the assumptions stated so far, the log likelihood of the change response pattern $\mathbf{y}_i$ is

$$L_{ch}(\theta_i) = \sum_j^n y_{ij} \ln(Pch_j(\theta_i)) + (1 - y_{ij}) \ln(1 - Pch_j(\theta_i)) . \quad (7)$$

By using the same rationale as Levine and Rubin's, the short-term consistency counterpart of the *lz* person-fit index, which we shall name *lz-ch*, is now defined as

$$l_{z-ch} = \frac{l_{o-ch} - E(l_{o-ch})}{[Var(l_{o-ch})]^{1/2}} \quad (8)$$

where *lo-ch* is the value of (7) evaluated at the ML estimate of *θ_i*, and the mean and variance are given respectively by

$$E(l_{0-ch}) = \sum_j^n Pch_j(\hat{\theta}_i) \ln(Pch_j(\hat{\theta}_i)) + (1 - Pch_j(\hat{\theta}_i)) \ln(1 - Pch_j(\hat{\theta}_i) \quad (9)$$

$$Var(l_{0-ch}) = \sum_j^n Pch_j(\hat{\theta}_i)(1 - Pch_j(\hat{\theta}_i)) \left[\ln \frac{Pch_j(\hat{\theta}_i)}{1 - Pch_j(\hat{\theta}_i)} \right]^2 \quad (10)$$

and $\hat{\theta}_i$ is the ML estimate of θ_i . Because the terms in (7) are independent for fixed *θ*, under the null hypothesis that all respondents follow the postulated model, the distribution of *lz-ch* is expected to approach the standard normal as the number of items increases. This expectation, however, is based on the same two strong assumptions that Drasgow et al. used when deriving the distribution of *lz*. First, the item parameters are assumed to be fixed and known. Second, the ML estimate of θ_i is expected to coincide with the 'true' value. To see these points, note that the *Pch(θ)* terms in (7) are treated as constants when deriving the mean and variance (9) and (10). And this treatment is only strictly correct when the two above assumptions are met.

The *lz-ch* statistic is intended to be used in a retest design in which the items are calibrated at Time 1, and the individuals are also scored at Time 1 by using the item estimates as fixed and known values. It is also assumed that the sample at Time 1 is large and representative enough for the assumption of known item parameters to be reasonable. However, the problem of using individual trait estimates in place of the true levels still remains. In the context of *lz,* some corrections have been proposed (e.g. Snijders, 2001). At the moment, however, we shall not still consider these corrections for reasons which are discussed below.

In summary, *lz-ch* is intended to be used in a situation in which a large and representative group of respondents is available at Time 1. The IRT model is calibrated in this group and the fit of the model is checked. Next, individual trait estimates are obtained at Time 1 by taking the item parameter estimates as fixed and known values.

Some or all of the individuals of the group are then retested after a short-term retest interval, and, for these individuals, the *lz-ch* values are obtained based on the item and person estimates at Time 1, and the observed responses at Time 1 and at Time 2. These values are used to flag those respondents that are potentially inconsistent in the short term.

As for possible cut scores, the standard normal distribution of *lz-ch* requires the fulfillment of some conditions that can never be fulfilled with real data. As well as the conditions discussed above (known item and person parameters), the model is assumed to be correct and the data is expected to contain no truly inconsistent respondents (i.e. the null hypothesis stated above). Models, however, are at best reasonable approximations, and the data is expected to contain a certain unknown proportion of inconsistent respondents. So, it seems unrealistic to expect normal distributions with real data, and, in the writer's opinion, it is inappropriate to suggest cut scores based on theoretical nominal levels. Future intensive research may make headway in this respect, including the corrections discussed above. For the moment it seems better to consider the standard normal distribution only as a useful reference.

Interpretation Issues and Further Extensions

Inspection of equations (7) to (10) shows that *lz-ch* values are negative when the respondent changes his/her item responses more often than expected. On the other hand, values are positive when the respondent tends to duplicate the responses given at Time 1 more often than would be expected given the IRT model and his/her trait estimate. This result suggests than in the *lz-ch* case both tails of the distribution should be examined, as large and positive values would probably indicate retest effects of the type discussed above. Furthermore, examination of the *lz-ch* distribution over respondents might also be useful for assessing retest effects. If the retest interval is too short for avoiding memory effects, the respondents as a group would tend to behave more consistently than expected. As a result, the mean of *lz-ch* in this group would be expected to shift toward positive values.

As mentioned above, one of the aims of *lz-ch* is to provide more information about consistency than can be obtained with *lz*. In principle, it seems clear that both indices assess different aspects of consistency. The *lz* index assesses consistency on a same-time/different-item basis, whereas *lz-ch* mainly assesses consistency on a different-time/same-item basis (although the consistency contribution is then summed over items). In spite of this distinction, however, if the assumptions on which both indices are based are fulfilled, then *lz* and *lz-ch* are probabilistically related. To see this point, consider a single item. If the response to this item at Time 1 is consistent, then *lz* will be positive (i.e. above the mean). If it is, according to the assumptions, the probability of also obtaining a positive *lz-ch* value is greater than the probability of obtaining a negative value. On the other hand, if the response at Time 1 is inconsistent, *lz* will be negative. And the probability of also obtaining a negative *lz-ch* value is greater than the probability of obtaining a positive value. From these results it follows that in a group of consistent respondents, *lz* and *lz-ch* are expected to be positively correlated. This result is theoretically sound. It seems reasonable to expect that those respondents who are the most consistent when responding in a single test situation also tend to be the most consistent under repetition.

Indeed, the index proposed here will be most useful when the respondent behaves inconsistently and there are discrepancies between *lz* and *lz-ch*. As a first example, consider a situation reported in the literature (Jorm, Duncan-Jones & Scott, 1989). The individual responds consistently at Time 1, but at Time 2 his/her motivation decreases and he/she responds more randomly or mechanically. In this case *lz* would probably indicate consistency, but *lz-ch* would not. As a second example consider the theta-shift model (Zickar & Drasgow, 1996). Assume that the administration conditions are neutral on the first occasion but that there is a strong motivation for faking good on the second occasion. Assume further that, under this pressure, the individual responds as if his/her trait level is more adapted than it really is. In this case *lz* would probability indicate consistency, perhaps even when it is computed on both occasions. However, because the *lz-ch* values are obtained based on the person estimates at the first occasion, *lz-ch* would be expected to flag this respondent as inconsistent given the temporary changes in his/her trait level.

While the information obtained with *lz-ch* seems useful, once a response pattern has been detected as potentially inconsistent, it also seems to be useful to collect more information about the specific item responses in which the inconsistent responding is more pronounced. Information of this type can be obtained by using individual item statistics. In particular, I propose to assess inconsistency at the individual item level by adapting the standardized residual statistic proposed by Wright (Wright & Stone, 1979) in the context of the Rasch model. In the framework used here (see equations 5 and 6), the standardized change residual for respondent *i* on item *j*, denoted by Zch_{ij} is obtained as

$$Z_{chij} = \frac{y_{ij} - Pch_j(\hat{\theta}_i)}{\left[Pch_j(\hat{\theta}_i)(1 - Pch_j(\hat{\theta}_i))\right]^{1/2}} \, . \qquad (11)$$

The Zch_{ij} measures the item *j* response contribution to the short-term inconsistency of individual *i*. So, a large positive value would be obtained if the individual changed a response to an item in which a change was very unlikely given the model and trait estimate. The aim of this paper is not to derive the specific relations between (8) and (11). However, we note that *lz-ch* in (8) can be re-expressed as a weighted sum of the discrepancy terms in the numerator of (11) (e.g. Snijders, 2001). So, it seems clear that the individual item inconsistencies directly contribute to the global inconsistency as measured by *lz-ch.*

As for possible cut scores, it should be stressed that the scaling in (11) (i.e. zero mean and unit variance) is only made so as to make the residuals more interpretable. However, it cannot be said that the Zch_{ij} values are distributed as a standard normal variable because, in fact, Zch_{ij} is simply a transformed binary variable. So, no attempt is made here to provide critical values based on exact probabilities. What is expected, however, is that the Zch_{ij} values will be regular enough to identify outlying responses reliably, and this point is considered in the illustrative example below. Finally, as a reference, we may consider that a Zch_{ij} value of +3 would be obtained if a respondent changed a response to an item for which the expected probability of change was as low as 0.10.

An Illustrative Example

The statistics and procedure discussed so far are illustrated using data collected by the author in personality research. A 60-item Neuroticism scale was administered twice in the same conditions to a sample of undergraduate students with a 4-week retest interval. The analyses that follow were based on the 436 respondents who were present at both administrations.

As discussed above, items were calibrated and respondents were scored using the Time 1 data. The items of the scale were dominance-based, of the type which are generally well fitted by the two-parameter IRT model (2PM, see e.g. Ferrando, 1994). Items were calibrated according to the 2PM in the normal-ogive metric by using BILOG MG-3 (Zimowski et al. 2003), and the unidimensionality assumption and the global model-data fit were assessed with NOHARM (Fraser & McDonald, 1988). The fit of the model, both at the global level and at the individual-item level was reasonably good (details of the goodness-of-fit results are available from the author). The item locations ranged from -3.12 to 4.15, with a mean of 0.75. The average of the item discriminations was 0.57.

The item parameter estimates obtained in the calibration stage were used in a real-parameter simulation study, which was parallel to the main empirical study. This parallel study simulated the responses of 436 individuals at two points of time to a test that behaved according to the 2PM, with item parameter values equal to those obtained in the calibration of the empirical data. In the simulated data all of the assumptions used in the derivation of *lz-ch* were met: The model was totally correct, the item responses were locally independent under repetition, and the sample did not contain truly inconsistent respondents. So, the simulated results were used to check the predictions about the behavior of *lz-ch* and to assess the discrepancies with respect to the real-data results.

Figure 1 shows the distribution of *lz-ch* with the simulated data (thick dashed curve) and with the empirical data (thick solid curve). The distributions were obtained by using Gaussian kernel density estimation (Silverman, 1986), which essentially provides smoothed improved histograms.

The distribution based on the simulated data agreed quite well with the expected standard normal distribution. The mean and the standard deviations were 0.05 and 0.95. The adherence of the *lz-ch* values to the corresponding theoretical distribution was further assessed by using the Kolmogorov-Smirnov statistic. The Kolmogorov distance was 0.05, with an associated probability of 0.13.

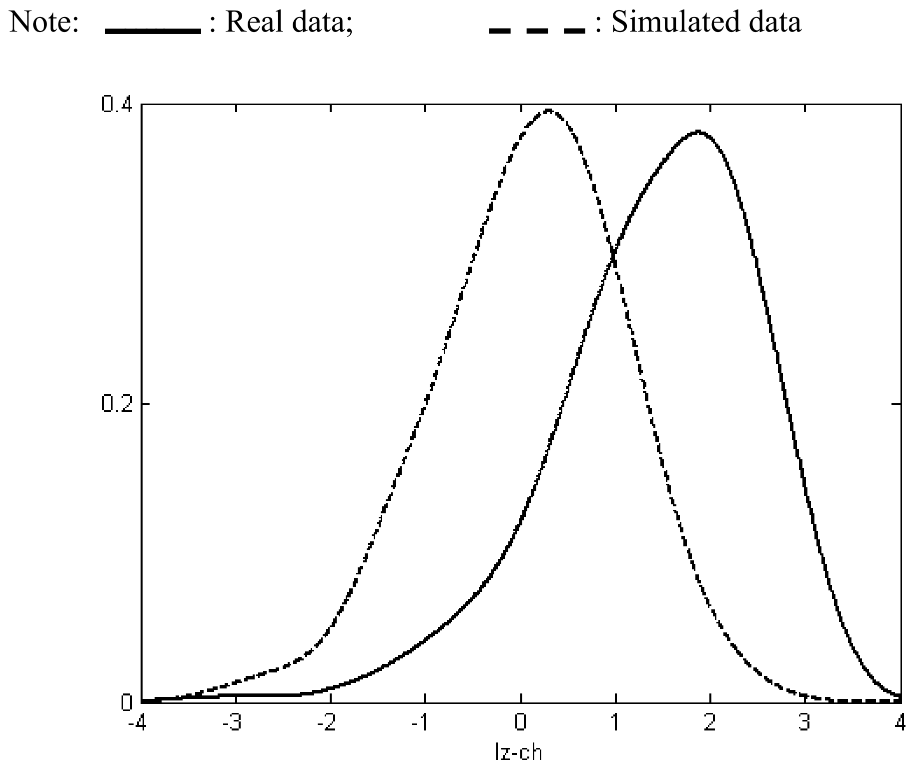

Figure 1. Distribution of *lz-ch* with the real data and the simulated data.

At first sight, the difference between the curves in figure 1 does not seem to be too great. However, the distribution based on real data, (a) is clearly shifted to the right (the mean was 1.38), (b) is somewhat asymmetrical with a heavier left tail, and (c) tends to have more dispersion than the standard normal distribution (the standard deviation was 1.17). Previous studies with *lz* show that when individual trait estimates are used in place of the true levels the variance usually decreases (Snijders, 2001, van Krimpen-Stoop & Meijer, 1999). So, the 'true' effect (c) might be even larger.

Results (b) and (c) are to be expected. The distribution of *lz-ch* is obtained by assuming that there are no inconsistent respondents in the group which is assessed. However, real data is expected to contain a certain, unknown proportion of inconsistent respondents. These are precisely the ones we wish to identify and who are expected to be found in the tails (particularly the left tail) of the distributions, thus increasing the dispersion.

As discussed above, result (a) would be expected if respondents tended to behave more consistently as a group than the model predicts. To further assess this issue, the expected number of response changes was obtained for each respondent by evaluating (5) at the estimated trait level, and summing the values over the 60 items. The values obtained were then compared to the observed number of response changes. The median of the expected number of changes over respondents was 20, whereas the median of the observed values was 10. The respondents clearly tended to behave more consistently in the short term than the model and the trait estimates would lead us to expect. Overall, these results suggest that the 4-week interval was insufficient to avoid retest (probably memory) effects.

The next factor to be assessed was the extent to which the proposed standardization (8) achieved independence from the trait levels estimated at Time 1. For the simulated data, the product-moment correlation between the θ estimates and *lz-ch* was -0.04. For the real data it was 0.08. Inspection of the scatterplots did not reveal a trend of any sort. So, the initial results suggest that the values of *lz-ch* are essentially independent from the trait levels. Indeed, far more research is needed on this issue.

We turn now to the relations between *lz-ch* and the *lz* values estimated at Time 1. For the simulated data the scatterplot suggested a linear relation without outliers, and the correlation was 0.52. This result agrees with the theoretical expectations discussed above. As for the real data, the scatterplot is shown in figure 2

Unlike the simulated case, figure 2 reveals the presence of outliers that affect the product-moment correlation, which in this case was only 0.32. These outliers are marked in figure 2 and are presumably respondents who behaved consistently at Time 1 (acceptable *lz* values) but inconsistently in the short-term. As discussed above, these are precisely the most interesting cases.

The *lz-ch* values for the individuals flagged in figure 2 were -5.67 (respondent 56), -3.36 (respondent 88) and -7.14 (respondent 357). These high negative values suggest that these respondents changed responses at Time 2 more often than expected. To further assess the type of inconsistency that was potentially detected by *lz-ch*, the standardized change residuals (11) were assessed for each item. For respondents 56 and 88, it was found that inconsistencies were located in a small group of items (15, 17, and 34 for respondent 56, and 28, 41, 46 and 51 for respondent 88). For these items, the Zch_{ij} values were far above 4 in all cases. In contrast, the *Zch*$_{ij}$ values for respondent 357 were systematically high (around 2) for most items but without the extremely high values obtained in the two previous cases. In applied research, further information should be collected to determine the cause of the unexpected changes of respondents 56 and 88. The behavior of respondent 357, on the other hand, rather suggests some type of systematic change in the trait levels.

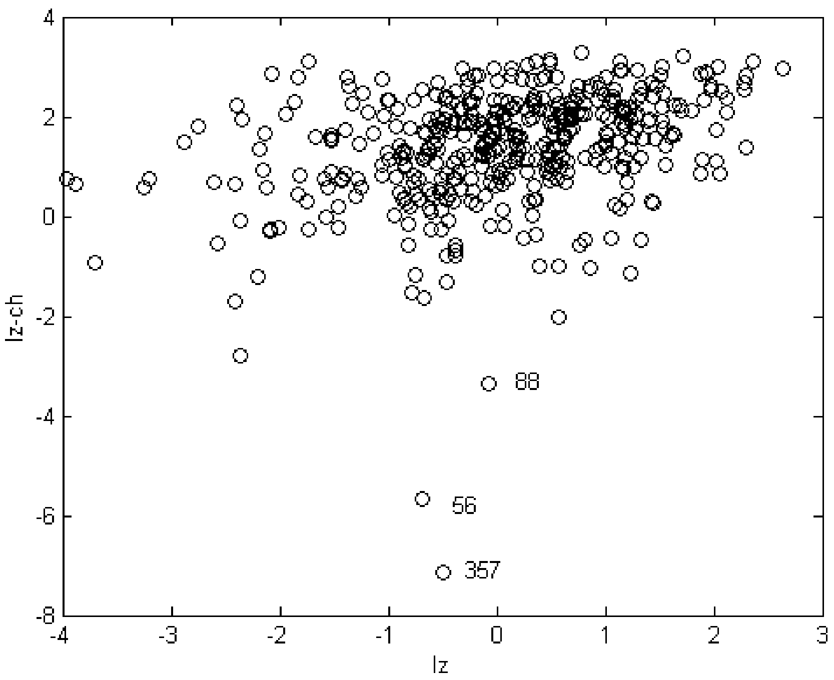

Figure 2. Relation between *lz* at Time 1 and *lz-ch*. Real data.

DISCUSSION

In personality measurement, short-term consistency is an important topic that has been the object of considerable research. Earlier studies used a classical test theory framework, whereas most modern research tends to be based on IRT. In both cases, however, the focus has generally been on the item characteristics that tend to elicit inconsistent responding (e.g. Ferrando et al. 2001). Furthermore, the respondents have only been studied at the group level, not at the individual level (e.g. Fischer, 1995). As far as the writer knows, no procedures at the individual level of the type discussed here have been proposed so far.

The present procedure is based on a likelihood-type index that can be considered as the short-term counterpart of the well known *lz* statistic. So, *lz-ch* can be assumed to have the same advantages and shortcomings as *lz*. Its advantages are that it has a clear rationale, is easy to interpret and is expected to have a relatively good power for detecting inconsistent respondents. Its disadvantages are that it only adheres to the standard normal reference distribution under strong assumptions that can never be fulfilled with real data. So, at present it does not seem appropriate to use the statistic for conducting strict tests of fit; rather it can be used at the descriptive level as a useful screening tool. Once a respondent has been flagged as potentially inconsistent, the standardized change residuals also proposed in this paper can be used to further assess the type of inconsistency detected by the global index, and perhaps to explain the reasons behind the inconsistencies. Further extensions can also be considered. One reviewer suggested that the individual estimates obtained separately at Time 1 and at Time 2 should be used for assessing specific sources of inconsistency. For example, for the theta-shift case, *lz-ch* could be modified so that it would take into account the estimated individual change. If so, the statistic would detect inconsistency beyond that due to the temporal change in the trait level.

The results obtained in the empirical study supported the expectations about the behavior of *lz-ch* . In the simulated data, based on a 60-item test with moderate discrimination and a wide range of item difficulties, the distribution of *lz-ch* approached the theoretical reference distribution quite well. In fact the departure from this distribution was non-significant according to the Kolmogorov-Smirnov test. Furthermore, the values of the statistic were virtually independent from the estimated trait level. With the real data, the departure from the reference distribution was clear. However, the discrepancies could be reasonably explained, and they provided useful information.

Finally, the main purpose of this paper is to provide a useful tool for applied researchers. In this respect, the procedure is relatively simple, and the proposed statistics can be easily programmed. At present, the writer is developing a user-friendly program that will be available at no cost.

RESUMEN

Evaluación de la consistencia individual a corto plazo mediante estadísticos basados en la TRI. Se propone un procedimiento, basado en un estadístico global para evaluar la consistencia intra-individual en un diseño retest basado en un intervalo de retest corto. El procedimiento se desarrolla en el marco de los modelos paramétricos de TRI, y el estadístico, que se basa en la función de verosimilitud, puede ser considerado como una extensión del estadístico *lz*. Se discute la justificación para el uso e interpretación del estadístico y se propone además un estadístico estandarizado a nivel de ítem para obtener claves acerca de las posibles causas de la inconsistencia detectada mediante el índice global. El procedimiento se ilustra con un ejemplo real y una simulación paralela en el dominio de la personalidad.

REFERENCES

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 and R.C. Johnson (eds.) *Functional Psychological Testing* (pp 54-78). New York: Brunner/Mazel.

Conley, J.J. (1984). The hierarchy of consistency: a review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences, 5,* 11-25.

Drasgow, F., Levine, M.V. & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Ferrando, P.J. & Chico, E. (2001).Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61,* 997-1012.

Ferrando, P.J.(1994). Fitting item response models to the EPI-A impulsivity subscale. *Educational and Psychological Measurement. 54,* 118-127.

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.

Fischer, G.H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459-487.

Fiske, D.W. & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52,* 217-250.

Fraser, C. & McDonald, R.P. (1988). NOHARM: least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Jorm, A.F., Duncan-Jones, P. & Scout, R. (1989). An analysis of the re-test artifact in longitudinal studies of psychiatric symptoms and personality. *Psychological Medicine, 19,* 487-493.

Knowles, E.S. (1988). Item context effects on personality scales: measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Levine, M.V. & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1,* 477-482.

Meijer, R.R.& Sijtsma, K. (1995). Detection of aberrant item scores patterns: A review and new developments. *Applied Measurement in Education. 8*, 261-272.

Meijer, R.R.& Sijtsma, K. (2001). Methodology review: Evaluationg person fit*. Applied Psychological Measurement. 25,* 107-135.

Meijer, R.R., Egberink, I.J.K., Emons, W.H.M. & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment, 90,* 1-14.

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

Reise , S.P. & Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S.P. & Flannery, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9-26.

Schmidt, F.L. & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Schuldberg, D. (1990). Varieties of inconsistency across test occasions: Effects of computerized test administration and repeated testing. *Journal of Personality Assessment, 55*, 168-182.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* London: Chapman & Hall.

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331-342.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 622-663.

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (ed.) *Educational Measurement* (pp. 560-619). Washington: American Council on Education.

van Krimpen-Stoop, E.M.L.A. & Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests*. Applied Psychological Measurement, 23*, 327-345.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38*, 319-350.

Wright, B. & Stone, C.H. (1979). *Best test design.* Chicago: Mesa press.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using apropriateness measurement. *Applied Psychological Measurement, 20,* 71-87.

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: 10 March 2009; accepted: 14 April 2009)