An Item Response Theory Analysis of Response Stability in Personality Measurement

Pere J. Ferrando and Urbano Lorenzo, 'Rovira i Virgili' University

Gabriel Molina, University of Valencia

An item response theory model of response stability is developed, based on the local independence principle. The model predicts response changes under repeated administrations of the same instrument using item and examinee parameter estimates as predictors. Real data were used to assess how the model functioned. Results indicated that the model predictions were approximately fulfilled. Limitations of the model and the empirical study are discussed. *Index terms: item response models and personality, local independence principle, personality measurement, test-retest stability.*

One weakness of personality tests is the marked instability of item responses under repeated administrations (Angleitner, John & Löhr, 1986; Fiske & Butler, 1963; Goldberg, 1963). The instability depends on the instrument used, the item response format, and on the test-retest interval. For binary items, the average percentage of examinees who change their responses in a typical personality item for a test-retest interval of 3 to 4 weeks is estimated to be 20 to 25% (Angleitner et al., 1986; Goldberg, 1963, 1978).

These results appear to contradict the data on the stability of personality scores for retest intervals of many years (Costa & McCrae, 1985; Smith, 1992). However, these results are based on total test scores—usually simple sums of item scores. In classical test theory (CTT), single items are unreliable, but linear composites of these items can be highly reliable (Gulliksen, 1950). Yet response instability at the item level is compatible with acceptable test-retest stability of total scores for long tests. For example, Schuerger, Zarella & Hotz (1989) noted that test-retest reliability can be predicted by the number of items and their homogeneity. At the item level, however, the problem remains.

Attempts to explain the instability of personality items have been based on item difficulty and ambiguity (Rogers, 1973). Fricke (1957) and Hanley (1962) suggested that items with extreme endorsement frequency were relatively easy to respond to, while items with intermediate endorsement frequency (controversial items) were difficult. Because most examinees were expected to guess at the answer to controversial items, it was predicted that these items would be less stable under repeated measurement.

Goldberg (1963) suggested that unstable items were ambiguous items. The meaning of the ambiguous items was open to various interpretations. The items were more likely to elicit response changes in a test-retest situation. Nowakowska (1983) noted that explanations for both the difficulty and ambiguity could be formulated in the framework of Thurstone's (1927) model of comparative judgment.

Several empirical studies analyzed the relations between the characteristics of items and examinees and the stability of item responses. They were all based on CTT and were mainly descriptive. For item characteristics, results can be summarized as follows:

1. The relationship between proportion of endorsement and stability (proportion of response change from the first to the second administration) is curvilinear (U-shaped). In accordance with predictions by Fricke (1957) and Hanley (1962), the extreme items were the most stable (Angleitner et al., 1986; Goldberg, 1963, 1978; Wiggins & Goldberg, 1965).
2. The relationship between the discrimination index (usually the point-biserial item-total correlation) and item stability is unclear. Some studies found nonsignificant correlations (Jones & Goldberg, 1967; Turner & Fiske, 1968), others found weak but significant positive correlations (i.e., more discriminating items were more stable; Angleitner et al., 1986), and earlier studies (reviewed in Goldberg, 1963) found that the most stable items were the least discriminating. This last result has been called the "psychometric paradox" (Goldberg, 1963; Nowakowska, 1983).

Nowakowska (1983) presented a mathematical model, based on CTT, relating item parameters to the probability that the answer to the item will change. The model leads to predictions that can be tested using real data. However, the model has not been tested empirically. Mitra & Fiske (1956) predicted that maximum change was expected at the "chance" score (the total score midpoint, in the case of binary personality items). Their empirical study found a curvilinear relationship between the number of items on which responses were changed on the second occasion and total test scores. The curve reached its peak at approximately the chance score, as predicted.

Purpose

Nowakowska's (1983) model is further developed within the framework of item response theory (IRT). This leads to a model for predicting item response change that relates item and test instability to the estimated values of item and examinee parameters. The model then is tested using real data. The model could be of interest because it could (1) help in the substantive interpretation of item parameters in personality measurement; and (2) be useful in the design and construction of a test, because it allows the researcher to predict, to some extent, the stability of a given item from the estimated values of that item's parameters and the stability of the responses of an examinee, given his/her estimated trait level.

Prediction of Change Model

The model analyzes Type 1, intra-individual variability (Fiske & Rice, 1955; Goldberg, 1978), which is the difference between an examinee's responses at two points in time when (1) the examinee is exposed to the same item on both occasions, and (2) the overall situation in which the responses are made is the same.

Consider a set of n binary items that measures a personality trait θ. Assume that they are all keyed in the same direction, so that a score of 1 corresponds to a higher level of θ.

Local independence is assumed to hold for repeated measurements of the same item. In particular, the conditional distributions of the responses to the same item in two repeated administrations are assumed to be independent of each other. This assumption is reasonable when a stable personality trait is measured and the test-retest interval is long enough to avoid memory or other retest effects.

Let $P_j(\theta_i)$ be the conditional probability of a score of 1 on item j, given a fixed value of θ. Under the above assumption, the conditional probability of a response change in this item is

$$Pch_j|\theta_i = 2P_j(\theta_i)\left[1 - P_j(\theta_i)\right] \; . \tag{1}$$

For example, consider a sample of randomly selected examinees, all of whom have the same θ of $-.5$. Suppose that the probability of endorsing a given item is $P = .2$ (the probability of not endorsing it is .8). If local independence holds, and the item parameters and θs do not change, the probability that an examinee will endorse this item the first time and not the second time is $.2 \times .8$—i.e., $P(\theta)[1 - P(\theta)]$—and the probability of not endorsing the first time and endorsing it the second is $.8 \times .2$. Thus, because the examinee can change the response in two ways, the total probability of change is $2 \times .2 \times .8$, as indicated in Equation 1. The unconditional probability of change in the item is then

$$Pch_j = 2 \int_{\theta} P_j(\theta)\left[1 - P_j(\theta)\right] f(\theta) d\theta \ . \tag{2}$$

Similar equations are found in Nowakowska (1983). The difference is that she used a CTT approach based on the true score instead of θ.

Using Equations 1 and 2, it is possible to predict item and test response changes when an item response model is specified for $P(\theta)$ and a density function is specified for θ. Because different choices of $P(\theta)$ and $f(\theta)$ lead to different predictions, it is important to select functions that are realistic and appropriate in the context of personality measurement.

The most appropriate model in the personality domain is the two-parameter logistic model (2PLM; Finch & West, 1997; Reise & Waller, 1990; Waller, Tellegen, McDonald, & Lykken, 1996), because the guessing parameter is assumed to be unimportant in this context. Therefore,

$$P_j(\theta_i) = \frac{\exp\left[Da_j(\theta_i - b_j)\right]}{1 + \exp\left[Da_j(\theta_i - b_j)\right]} \ , \tag{3}$$

where

a_j is the item discrimination parameter,
b_j is the item location parameter (frequently referred to as item difficulty), and
$D = 1.702$ is a scaling factor that brings the logistic model into close agreement with the normal ogive model.

The $f(\theta)$ distribution can either be specified "a priori" (the most common being standard normal), or estimated from the data. If the standard normal distribution is assumed,

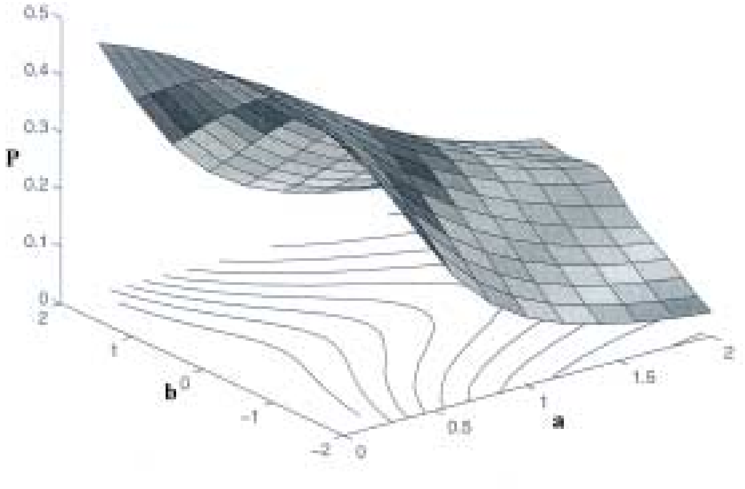
$$f(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta^2\right) \ . \tag{4}$$

Using Equations 2–4, the expected probability of response change in a given item was computed for an item with $a = (0.0, .2)$ and $b = (-2.0, 2.0)$. Using 20 points of Gaussian quadrature (Stroud & Secrest, 1966), the integral in Equation 2 was evaluated. The response change surface is shown in Figure 1. As Figure 1 shows, the predicted relationship between b and the proportion of change is an inverted-U curve and the predicted relationship between a and the proportion of change is negative. Both indices interact when predicting the proportion of change, so if they were used separately (e.g., by computing the correlation between a and the proportion of change), the resulting prediction would be unclear.

The expected proportion of endorsement for a given item is

$$P_j = \int_{\theta} P_j(\theta) f(\theta) d\theta \ . \tag{5}$$

Figure 1
Response Change Surface for a Two-Parameter Item
With $a = (0.2, 2.0)$ and $b = (-2.0, 2.0)$



Using Equations 2 and 5, relationships between the proportion of endorsement and the proportion of response change can be predicted. For fixed as, the relationships are inverted-U curves, as was previously found.

According to Equation 1, the conditional probability of response change for a given item is maximal when $P(\theta) = .5$. If the 2PLM is correct for the data, the maximum proportion of response change is expected for examinees whose θs are equal to b (i.e., the level at which the item provides the most information about the person being measured; Lord, 1980). When $\theta = b$, examinees could feel that both responses are applicable. Therefore, they will most likely change their responses to the items about which they are unsure (Kuncel & Fiske, 1974; Mac Eaton & Fiske, 1971).

Let x be the number of test items for which responses changed on the second occasion. For fixed θ, the repeated measurement in each item can be considered a Bernouilli variable with two values: 0 (no change) and 1 (change) with probability $Pch|\theta$. If local independence holds, the conditional distribution of the number of items for which responses changed is a sum of independent Bernouilli variables with different probability values. This is a generalized binomial distribution (Kendall & Stuart, 1977, p. 134). The conditional mean of x is

$$\mathrm{E}(x|\theta) = \sum_{j=1}^{n} Pch_j|\theta = 2\sum_{j=1}^{n} P_j(\theta)\left[1 - P_j(\theta)\right] = 2n\left[\overline{P}(\theta)\left[1 - \overline{P}(\theta)\right] - \sigma^2_{p|\theta}\right] . \tag{6}$$

The variance of x for fixed θ is

$$\sigma^2(x|\theta) = \sum_{j=1}^{n}(Pch_j|\theta)\left[1 - (Pch_j|\theta)\right] . \tag{7}$$

In Equation 6, $\overline{P}(\theta)$ is the average of $P_j(\theta)$ taken over n items, and $\sigma^2_{p|\theta}$ is the variance of $P_j(\theta)$ for fixed θ taken over items. In practice, the expected values given by Equation 6 for different values of θ depend primarily on $\overline{P}(\theta)$. Therefore, the maximum change is expected at approximately the θ level at which the average of the conditional probabilities is .5.

The predictions made using Equation 6 are more specific than those made by Mitra & Fiske (1956), because the conditional probabilities are considered at each θ level. Mitra and Fiske considered only the examinees who randomly responded to all n items. For these examinees, the distribution of the total test scores was binomial with parameters n and $p = .5$. The expected mean score was $n/2$—the scale midpoint or the "chance" score. Mitra and Fiske further assumed that examinees with test scores near the chance score were more likely to have responded randomly, so they predicted the maximum change score at this level.

In the 2PLM, the test information function is (Lord, 1980)

$$I(\theta_i) = \sum_{j=1}^{n} D^2 a_j^2 P_j(\theta_i) \left[1 - P_j(\theta_i)\right] \ . \tag{8}$$

Comparing Equations 6 and 8, if all as are equal, the maximum number of expected item changes would be at the θ at which the test information function is maximal. The expected observed score (true score) is (Lord, 1980)

$$T(\theta_i) = \sum_{j=1}^{n} P_j(\theta_i) \ . \tag{9}$$

Equations 6 and 9 relate the expected observed scores to the expected number of response changes.

Method

Instrumentation and Participants

Sixty binary items were selected from the Neuroticism scales of the Maudsley Medical Questionnaire (Eysenck, 1952), the Eysenck Personality Inventory (Forms A and B; Eysenck & Eysenck, 1969) and the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1976). Angleitner et al. (1986) compared seven Anglo-American personality questionnaires and found that the Neuroticism scales were the most internally consistent and provided the highest parallel scales correlation. Because most of the present predictions refer to items, a relatively large pool of nonredundant items was used so that results were more stable and generalizable.

The 60-item instrument was administered twice in the same conditions and by the same examiners. The Time 1 sample included 625 undergraduate students, and the Time 2 sample had 527. The number of examinees present at both administrations was 432. All examinees were studying Psychology and Social Sciences at a university in Spain. Approximately 80% were female, with a mean age of 21. At Time 1 and Time 2, examinees completed questionnaires in classroom groups. The test-retest interval was 4 weeks in all cases.

Procedures

The prediction-of-change model was examined in a series of steps.

Step 1: Unidimensionality assessment and item calibration. The model described in Equations 1–9 is unidimensional, and the item parameters are assumed to be fixed and known. However, in any empirical study, the parameters must be estimated from the sample and the unidimensionality assumption should be tested. If data are multidimensional and a unidimensional model is used, the parameter estimates might be degraded (Way, Ansley, & Forsyth, 1988); this, in turn, might obscure the relationships predicted by the model.

Unidimensionality was assessed using two procedures: (1) NOHARM's harmonic analysis of the normal ogive model (Fraser & McDonald, 1988), and (2) Stout's T index of essential unidimensionality using DIMTEST (Stout, Douglas, Junker, & Roussos, 1993). NOHARM tests the usual factor-analytic definition of unidimensionality that all residual covariances are zero after the factor analysis; T tests a weaker hypothesis—for fixed trait levels, the average absolute inter-item covariance approaches zero as the test length increases.

In NOHARM, the item responses are assumed to conform to the unidimensional two-parameter normal ogive model. This model is approximated by a polynomial series on the basis of harmonic (Fourier) analysis. The item parameters are estimated with unweighted least squares by assuming that the latent trait is a random variable with a normal distribution. The use of unweighted least squares allows NOHARM to handle large datasets.

The assessment of unidimensionality in DIMTEST uses three steps.

1. The items are split into two short assessment subtests (AT1 and AT2), and a long partitioning subtest (PT).
2. The examinees are assigned to groups based on their PT scores.
3. T is computed using three estimates based on AT1: (1) the observed variance for the group, (2) the "unidimensional" variance estimate, and (3) the standard error of estimate. AT2 is used to correct T for bias. If unidimensionality holds, the observed variance will be similar to the unidimensional estimate, because the inter-item covariances will be small within each group. If the assumption of essential independence fails within subgroups, T will be large and the null hypothesis of essential unidimensionality can be rejected.

The indices used in NOHARM were the root mean squared residual (RMSR) covariances after fitting models of one and two factors, and the unweighted least squares version of Tanaka & Huba's (1985) γ goodness-of-fit index (McDonald & Mok, 1995). In NOHARM, unidimensionality was assessed separately in the Time 1 and Time 2 samples.

Stout's procedure works well in datasets of more than 25 items and 750 examinees (Stout, 1987) and loses power in sample sizes of approximately 500 examinees (Nandakumar, 1994). Therefore, only the Time 1 sample ($N = 625$) was tested using DIMTEST. The items in AT1, AT2, and PT were automatically selected by DIMTEST using a principal axis factor analysis of the tetrachoric correlation matrix based on 250 examinees; T was computed for 375 examinees.

Item and examinee parameters were estimated for both Time 1 and Time 2 data using BILOG 3 (Mislevy & Bock, 1990). Item parameters of the 2PLM in the normal metric ($D = 1.702$) were estimated using marginal maximum likelihood. Examinee θ was estimated using expected a posteriori (EAP). Item parameters were estimated separately in the Time 1 ($N = 625$) and Time 2 ($N = 527$) data by specifying a standard normal prior latent distribution and by estimating the latent distribution from the data using a finite number of points (Mislevy, 1984).

Step 2: Prediction of response changes based on item parameter estimates. A program written in MATLAB (1999) obtained the expected proportion of response change in each of the 60 items, using Equation 2 and the parameter estimates obtained from the Time 1 data. Item parameter estimates obtained by BILOG were used in Equation 3. The latent distribution of θ was approximated by a discrete distribution in 20 points (Mislevy, 1984). The relationship between the observed and predicted proportions of response change was assessed by linear regression.

Step 3: Prediction of changes based on examinee parameter estimates. A program written in MATLAB estimated the number of item responses that would change at Time 2 and the variance for fixed θ according to Equations 6 and 7. The program used the item parameter estimates obtained by BILOG. The relationship between the observed and expected number of response changes for all θ estimates was examined graphically.

Results and Additional Analyses

Testing the Unidimensionality Assumption and Item Calibration

Unidimensionality. The results of the assessment of fit are shown in Table 1. The one-factor model fit reasonably well according to NOHARM and DIMTEST ($T = 1.313$, $p = .092$). The NOHARM indices also indicated that the one-factor solution could not be improved by using a two-factor model, and that the fit was quite similar in the Time 1 and Time 2 data. Overall, the results indicated that the data were essentially unidimensional: there was a single dominant trait underlying the item responses.

Table 1
RMSR and γ Goodness-of-Fit (γ-GFI) Results From NOHARM for Time 1 ($N = 625$) and Time 2 ($N = 527$) Under the 1- and 2-Factor Models

Model and Index	Time 1	Time 2
1-Factor		
RMSR	.0121	.0127
γ-GFI	.906	.912
2-Factor		
RMSR	.0106	.0107
γ-GFI	.928	.937

Item parameters. The fit of the 2PLM was examined using the χ^2 statistic implemented in BILOG for long tests. Nonsignificant χ^2s were obtained for all 60 items, suggesting that the model was appropriate for the present data.

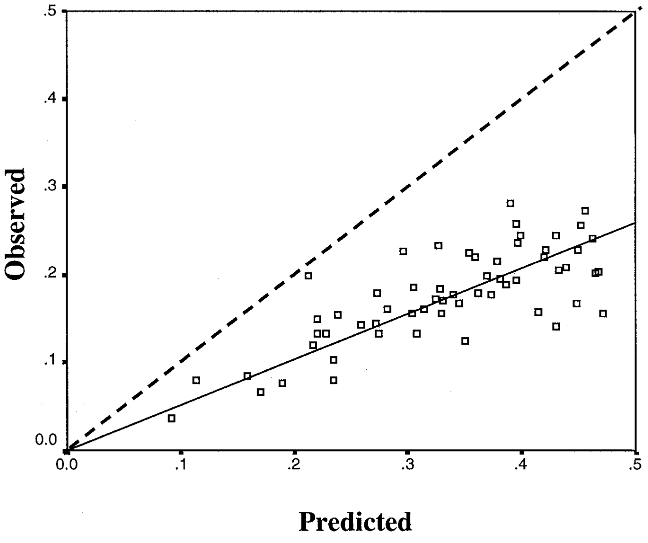
Correlations between the corresponding parameters estimated within the same sample with different priors were never smaller than .999. This suggests that in this case it was of no practical importance whether a normal prior distribution or a prior estimated from the data were used. For this reason, only results obtained from the estimated latent distributions were considered. The a estimates at Time 1 and Time 2 correlated .950, and the b estimates correlated .981. This suggests that both parameters were essentially invariant under the test-retest situation. Given these high correlations, only the estimates obtained in the Time 1 sample were used in all predictions.

Prediction of Response Changes Based on Item Parameter Estimates

The observed proportion of response change was computed for each of the 60 items. Results ranged from .036 to .282, with a mean of .176. These results agreed closely with those obtained with neuroticism scales of different Eysenck questionnaires by Angleitner et al. (1986).

Figure 2 plots the observed proportion of response change in the test-retest for each of 60 items against the proportion of response change predicted by the model from the item parameter estimates. The dashed line represents equality in the proportions. The solid line is the linear regression line through the origin with slope .518 (the intercept estimate of .03 was nonsignificant). The product-moment correlation (r) and the root mean squared deviation (RMSD) between the observed and predicted proportions were .764 and .0758, respectively.

Figure 2
Observed and Predicted Proportion of Response Change in 60 Items



These results indicate that the relationship between the observed and predicted proportions of response change in the items was proportional and rather strong. Second, for all items, the observed proportion of change was smaller than the predicted proportion.

The relationship between the item parameters and item stability was stronger than assumed by earlier studies [see Goldberg (1963) for a review]. The stability of the item responses can be reasonably predicted from the item parameter estimates. A perfect relationship, such as that in Figure 1, is unattainable for several reasons: (1) the model for predicting change is a population model, whereas in any empirical study, item parameter values, distribution of θ, and observed proportions of change are sample estimates; (2) the 2PLM is assumed to hold in the population, but is only an approximation in any sample; (3) the 2PLM assumes perfect unidimensionality, but there was only a single dominant dimension underlying the item responses; and (4) the model does not take into account possible individual changes in θ that might take place between testing.

Simulation Study

Method. A simulation study was implemented to assess whether lack of fit to the 2PLM was responsible for the results observed in the empirical data. The conditions of the study were selected to produce datasets that were as similar as possible to the conditions of the empirical study: (1) the data were essentially unidimensional, (2) the 2PLM was found to be approximately appropriate for all items, and (3) the estimated latent distribution was approximately normal with a slight positive skew.

Conditions 1 and 2 were simulated by specifying a two-dimensional 2PLM in the population (Reckase, 1997). Population item parameters were based on the parameter estimates obtained in the two-dimensional NOHARM solution for the empirical data. In this way, the population model was essentially unidimensional, with a first strong dominant dimension and a second minor dimension, and the item parameter values were similar to those of the empirical data.

Two two-dimensional θ distributions were considered: normal and positively skewed (χ^2 with 4 degrees of freedom). In both cases, θ vectors were specified to have a mean of 0.0 and a standard deviation of 1.0 in the population. Because the item parameters were taken from the NOHARM orthogonal solution, the population correlation between θ vectors was 0.0.

Data were generated using MATLAB's random number simulator. For the given parameters, the two-dimensional 2PLM was used to generate probability matrices with 400 persons × 60 items. The probability values were compared with two matrices with elements generated from a uniform distribution in the range 0.0 to 1.0. In each case, if the random number was less than or equal to the corresponding probability value, the item response was set to 1; otherwise, it was set to 0. The Time 1 responses resulted from the comparison to the first matrix and the Time 2 responses from the comparison to the second. Once the response matrices for Time 1 and Time 2 had been obtained, the remaining analyses were the same as in the empirical study (i.e., BILOG analyzed the data as if it were unidimensional and as if the latent prior were normal). 50 replications were used under the normal distribution condition and 50 under the positively skewed distribution condition.

Results. When θ was normally distributed, the correlations between observed and predicted proportions of change ranged from .939 to .976, with a mean of .956. The regression slopes through the origin ranged from .989 to 1.025, with a mean of 1.007. The RMSD ranged from .013 to .022, with a mean of .018. When θ was skewed, the correlations ranged from .924 to .972, with a mean of .955. The slopes ranged from 1.004 to 1.041, with a mean of 1.027. The RMSD ranged from .014 to .025, with a mean of .020. These results indicated that the conditions considered were not responsible for the constant bias found in the empirical study. The positively skewed distribution did not cause any important distortion compared with the estimations obtained when the distribution was normal.

An alternative explanation for the bias observed in the empirical study might be retest effects—memory and specific error effects. Retest effects mean that examinees tended to duplicate their former response; therefore, the responses ceased to be locally independent. Although the four-week interval used here was considered sufficient for avoiding memory effects (Cattell, 1986; Goldberg, 1978), some researchers believe that these effects might occur even with retest intervals of many years (Nunnally, 1970). Explanations based on specific error effects assume that there are specific factors that are the result of peculiarities of item content, and their influence on item scores is the same on both occasions. Schmidt & Hunter (1996) suggested that the effect of specific error was particularly large in personality trait measurement.

Follow-Up Empirical Study

Data. To study the plausibility of the explanations for the bias based on retest effects, a second empirical study was conducted using data from a Spanish adaptation of Dickman's (1990) Impulsivity Inventory (DII). This version of the DII consisted of 11 binary scored items designed to measure functional impulsivity, and 11 items designed to measure dysfunctional impulsivity. The DII was administered twice under the same conditions by the same examiner. The sample was composed of 182 undergraduate students; the number of examinees present at both administrations was 106. The test-retest interval was 10 weeks in all cases.

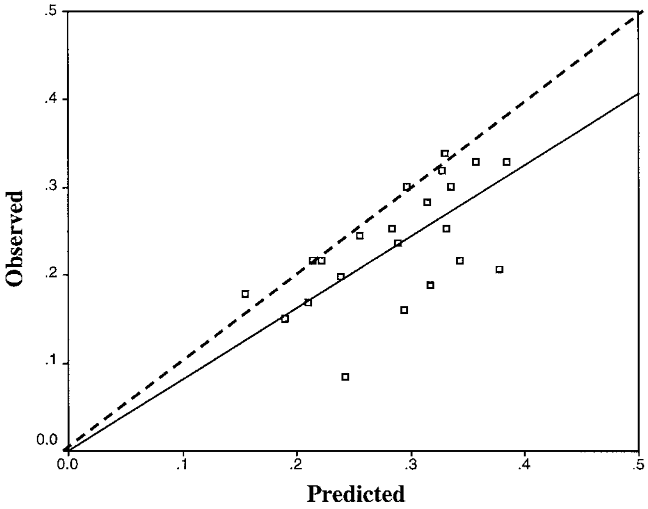
The two DII subscales were calibrated separately, but all 22 items were used together in the prediction so that the results were more stable. Because of the small sample size, only NOHARM was used to assess unidimensionality.

Results. The RMSR covariances and the γ goodness-of-fit index after fitting the one-factor model were .017 and .916, respectively, for the functional impulsivity subscale, and .014 and .938, respectively, for the dysfunctional impulsivity scale. Both scales were essentially unidimensional.

Item parameters were estimated using BILOG 3. However, because of the small sample size, no attempt was made to estimate the latent distribution from the data; a standard normal prior was specified instead. Inspection of the item parameter estimates revealed a narrow spread of b, which were almost all greater than 0.0.

Figure 3 plots the observed proportion of response change in the test-retest for each of 22 items against the proportion of response change predicted by the model from the item parameter estimates. The linear regression through the origin had a slope of .815. The product-moment correlation between the observed and predicted proportions was .623, and the RMSD was .045.

Figure 3
Observed and Predicted Proportion of Response Change in 22 Items



As the slope and RMSD indicate and Figure 3 illustrates, the consistent bias found in the first study was much less pronounced in the second study, which used a longer retest interval. Therefore, the explanation of bias based on memory effects appears to be plausible. However, both studies differ in many characteristics (different instruments, different number of items), and the reduction of bias cannot be unambiguously attributed only to the use of a longer retest interval.

The correlation between the observed and predicted proportions was smaller in the second study, but this was expected because: (1) item parameter estimation was less accurate due to the smaller sample size, (2) there were fewer items, and (3) there was restriction in the range of b. The model predicted reasonably well in the second dataset. The model's usefulness with two different instruments and with data collected in different samples suggests that it has some generalizability.

Prediction of Change Based on Examinee Parameter Estimates

For each of the 432 examinees in the first study, the number of items on which responses were changed on the second occasion was computed. The number of changes per examinee ranged from 1 to 24, with a mean of 10.42.

Figure 4 plots the observed (squares) and predicted (crosses) number of items with changed responses for each examinee against the EAP θ estimate. EAP estimates were rescaled so that the mean was 0 and the variance 1, as assumed in the estimated latent distribution. The band around the prediction was obtained by connecting the points marking off the 95% confidence intervals of the conditional distribution in Equations 6 and 7 at different θs. The confidence intervals were computed using the normal approximation to the conditional generalized binomial distribution.

Figure 4
Number of Test Response Changes and θ Estimates for 432 Examinees

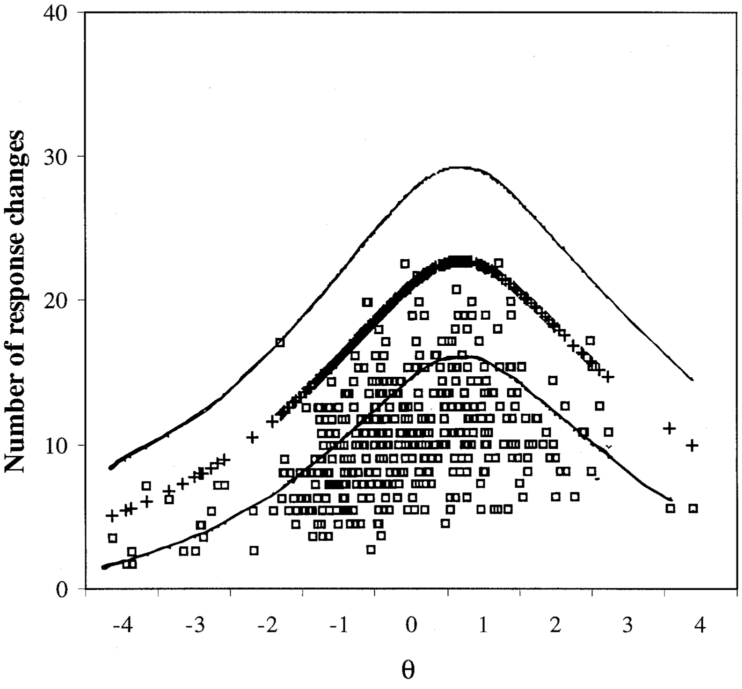


Figure 4 shows that (1) there was a relationship between $\hat{\theta}$ and the number of response changes in the retest, as predicted by the model (i.e., the maximum change was expected around the θ level at which the average of the conditional probabilities was .5); (2) the observations were more scattered than was predicted; and (3) the number of observed changes was systematically lower than the number predicted (as discussed above). The second result was expected because θs were not known population values, but were estimates from the response to only 60 items. The item parameter values were also estimates.

The model predicted that the maximum number of response changes was expected at $\theta = .62$. At this point, the average of the conditional probabilities was .486—quite close to .5 (see Equation 6). The maximum test information was estimated at $\theta = .75$. From Equations 6–8, the model predicted the maximum number of changes around those levels of θ at which the test was most informative. These were also the levels at which the dispersion of the number of responses changed was expected to be maximal.

Person-fit analysis. A person-fit statistic was computed for each examinee at Times 1 and 2. An examinee identified as misfitting at Time 1 was predicted to change his/her response more often at Time 2. θs are estimates, not known values, so if examinees with poorly estimated θs (high misfit) were eliminated, the estimates of the relationship between the number of changes and θ might be more accurate.

The standardized, log-likelihood, person-fit statistic l_z (Drasgow, Levine,& Williams, 1985) was the index used based on evaluations of person-fit statistics by Nering & Meijer (1998). Negative l_zs (usually less than -2.00) indicate that an examinee's response pattern is not consistent with his/her θ estimate. l_z was computed for each examinee at Times 1 and 2. In both cases, the distributions were smooth, unimodal, and slightly negatively skewed. The means and standard deviations of l_z scores were, respectively, .025 and 1.124 (Time 1), and .021 and 1.020 (Time 2; expectations were 0.0 and 1.0, respectively). The correlation between l_zs at Time 1 and Time 2 was .645. This correlation was higher than the split-half reliability estimates of the l_z scores reported in other personality studies (Reise & Waller, 1993).

The correlation between l_z at Time 1 and the number of response changes was $-.128$, which was statistically significant ($p < .01$) and in the expected direction (i.e., negative l_zs tended to be associated with a larger number of changes). However, the value was quite low. This could be due to there being a variety of response types that could give rise to nonfitting response vectors, but that would not necessarily lead to changes in response at Time 2 [e.g., deliberate faking or dissimulation (Birenbaum, 1986), language difficulties, response bias, or idiosyncratic personality trait structures (Reise & Waller, 1993)].

Of the 432 examinees, 24 (5.5%) had l_z below -2.00. Removing these examinees did not change the nonlinear relationship between the number of response changes and $\hat{\theta}$ (Figure 4). This result appears to agree with previous findings (e.g., Meijer, 1997).

Discussion and Conclusions

Two primary conclusions can be drawn from this study. First, if an appropriate IRT model is used, there is a clear relationship between the item parameter estimates and the proportion of item response change observed when the same instrument is administered on a second occasion. Second, there is a relationship between estimated θ and the number of test items on which responses changed in the retest. However, this relationship cannot be predicted as precisely as the relationship based on item parameters.

For a given item, the basic prediction of the model is that the maximum proportion of response change is expected for examinees whose θs are equal to the item difficulty, b. The b parameter can be considered a threshold point in the θ dimension in the sense that, for θs greater than b, the probability of endorsing an item is higher than the probability of not endorsing it, whereas for θs lower than b, the opposite is true. Uncertainty about the response, therefore, is maximal when $\theta = b$ and, if local independence is assumed, this is also the level at which the probability of change is greatest. This interpretation of b as a threshold is similar to the concept of response threshold in Jackson's (1986) model of personality responding.

Goldberg's (1963) model of item ambiguity considered not only a threshold point in θ, but also an "ambiguity band" or transition area around the threshold, so that examinees with θs in the ambiguity band were those who had the most difficulty responding to the item. This could be relevant to the present study. In particular, in the 2PLM, the item discrimination a is proportional to the slope of the item response function at $\theta = b$ [i.e., when $P(\theta) = .5$]. If an ambiguity band around b is considered, then a can be considered an index of the width of this band, because as a increases, $P(\theta)$ changes from .5 as θ moves from b in any direction.

If θ is distributed so that most examinees are concentrated in the central area and the frequencies decrease as they move toward the tails (e.g., in the normal distribution), then few examinees will have θs close to the bs of the extreme items. Therefore, items with extreme bs will tend to be the most stable under repetition. Nonextreme items, however, could also be stable if their as were large enough. As the model predicted and Figure 1 shows, the probability of change is expected to be maximal when b is 0.0. However, the larger the a, the narrower the region around 0.0 in which the probability of change can be high.

Costa & McCrae (1985) suggested that current studies provide stronger evidence of stability in the personality domain. This is partly because current psychometric instruments are much better than those used previously. This agrees with the previous discussion, because the quality of an item is given by its discrimination (Lord, 1980).

The expected relationship between a and the probability of change is conceptually more plausible than the psychometric paradox, which is considered to be a statistical artifact. Extreme items have low variances, so the item-total point-biserial correlations tend to be attenuated for these items. Because extreme items tend also to be the most stable, the stable items in some studies tended to have lower item-total correlations. However, the psychometric paradox does not disappear in this IRT analysis—it simply adopts another form. Item instability is predicted to be maximal around the θ at which the item discrimination power and score instability are maximal (i.e., the test information function is maximal).

Results of the present study differ from previous studies, because empirical studies were typically based on small samples and large pools of items that were often collected from different instruments and measured different traits. They were not based on a psychometric model and, therefore, tested no assumption or type of fit. The relationships between item parameters and the proportion of item response change previously were studied separately, often using only linear correlations. In contrast, this study was based on a specific mathematical model that considered a and b jointly. It also was founded on a series of assumptions (unidimensionality, appropriateness of the item response model) that could be statistically tested. For the examinee-based predictions, those based on examinee parameters were much more specific than those made by Mitra & Fiske (1956), who only considered examinees with random responses.

The present model has clear limitations. It is extremely simple and does not take into account several factors that can modify the prediction, including: (1) retest effects due to memory or specific error (Nunnally, 1970; Schmidt & Hunter, 1996); (2) changes in consistency due to experience with the test (Fiske, 1966; Knowles, 1988); and (3) individual changes in θ, such as tremors (moment-to-moment shifts) and swells (short-term mood swings) that take place between testing (Lumsden, 1977). However, these factors are very difficult to model mathematically. Also, the model here led to reasonable predictions, despite its simplicity.

The results of the empirical study can be generalized only to a certain extent. Two pools of items from two different personality instruments were used. However, it would be of interest to determine the extent to which the predictions are general to personality instruments, or whether they are limited to certain personality constructs or questionnaires.

References

Angleitner, A., J., O. P., & Löhr, F. J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner and J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 61–107). Berlin, Germany: Springer-Verlag.

Birenbaum, M. (1986). Effects of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167–174.

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.

Costa, P. T., & McCrae, R. R. (1985). Concurrent validation after 20 years: The implications of personality stability for its assessment. In J. N. Butcher and C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 4., pp. 31–54). Hillsdale NJ: Erlbaum.

Dickman, S. J. (1990). Functional and dysfunctional impulsivity: Personality and cognitive correlates. *Journal of Personality and Social Psychology, 58,* 95–102.

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.

Eysenck, H. J. (1952). *The scientific study of personality.* London: Routledge.

Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement.* London: Routledge.

Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality.* New York: Crane.

Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality, 31,* 439–485.

Fiske, D. W. (1966). Some hypotheses concerning test adequacy. *Educational and Psychological Measurement, 26,* 69–88.

Fiske, D. W., & Butler, J. M. (1963). The experimental conditions for measuring individual differences. *Educational and Psychological Measurement, 23,* 249–266.

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.

Fricke, B. G. (1957). A response bias (B) scale for the MMPI. *Journal of Counseling Psychology, 4,* 149–153.

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

Goldberg, L. R. (1978). The reliability of reliability: The generality and correlates of intra-individual consistency in responses to structured personality inventories. *Applied Psychological Measurement, 2,* 269–291.

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

Hanley, C. (1962). The "difficulty" of a personality inventory item. *Educational and Psychological Measurement, 22,* 577–584.

Jackson, D. N. (1986). The process of responding in personality assessment. In A. Angleitner and J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 123–142). Berlin, Germany: Springer-Verlag.

Jones, R. R., & Goldberg, L. R. (1967). Interrelationships among personality scale parameters: Item response stability and scale reliability. *Educational and Psychological Measurement, 27,* 323–333.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1). London: Griffin.

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

Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement, 34,* 743–755.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

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

Mac Eaton, A., & Fiske, D. W. (1971). Item stability as related to implicit set and subject-item distance. *Journal of Consulting and Clinical Psychology, 37,* 259–266.

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

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23–40.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21,* 99–113.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3 item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Mitra, S. K., & Fiske, D. W. (1956). Intra-individual variability as related to test score and item. *Educational and Psychological Measurement, 16,* 3–12.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses—comparison of different approaches. *Journal of Educational Measurement, 31,* 17–35.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22,* 53–69.

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

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Reckase, M. D. (1997). A linear logistic multidimensional item response model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). Berlin, Germany: Springer-Verlag.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14,* 45–58.

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

Rogers, T. B. (1973). Toward a definition of the difficulty of a personality item. *Psychological Reports, 33,* 159–166.

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

Schuerger, J. M., Zarella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 56,* 777–783.

Smith, D. D. (1992). Longitudinal stability of personality. *Psychological Reports, 70,* 483–498.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST manual.* Urbana Il: Department of Statistics, University of Illinois.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs NJ: Prentice-Hall.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38,* 197–201.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Turner, C. B., & Fiske, D. W. (1968). Item quality and appropriateness of response processes. *Educational and Psychological Measurement, 28,* 297–315.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and validation of a negative emotionality scale. *Journal of Personality, 64,* 545–576.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239–252.

Wiggins, J. S., & Goldberg, L. R. (1965). Interrelationships among MMPI item characteristics. *Educational and Psychological Measurement, 25,* 381–397.

Acknowledgments

The authors thank Eliseo Chico of ‘Rovira i Virgili’ University for providing the data used in the follow-up empirical study.

Author’s Address

Send requests for reprints or further information to Pere Joan Ferrando, Universidad ‘Rovira i Virgili’, Facultad de Psicologia, Carretera Valls s/n, 43007 Tarragona, Spain. Email: pjfp@fcep.urv.es.