*Psicológica (2006), 27*, **137-148.**

# Person-item distance and response time: An empirical study in personality measurement

Pere J. Ferrando*

'Rovira i Virgili' University (Spain)

This study assessed the hypothesis that the response time to an item increases as the positions of the item and the respondent on the continuum of the trait that is measured draw closer together. This hypothesis has previously been stated by several authors, but so far it does not seem to have been empirically assessed in a rigorous way. A computerized version of a 22-item two-scale personality questionnaire was administered to a sample of 286 respondents. The item responses were fitted using the two-parameter IRT model and a person-item distance measure was derived. Product-moment correlations between the log-response times and the person-item distances were obtained over respondents within each item. In both scales all the correlations were negative, as expected from the theory. However, most of the correlations (effect sizes) were small. The potential usefulness of the results for personality measurement is discussed.

The so-called 'distance-difficulty (DD) hypothesis' has been put forward repeatedly as one of the mechanisms involved in the process of responding to a personality item (e.g. Eisenberg and Wesman, 1941; Kuncel, 1977; Kuncel and Fiske, 1974; Tyler, 1968). It is intended for binary items that measure a single trait, and, in general terms, states that the difficulty of responding to an item increases as the trait level of the respondent approaches the location of the item on the continuum of the trait that is measured. The DD hypothesis was developed as an analogy with a well-known psychophysical result: that the uncertainty of responding to an

* **Acknowledgments:** This research was supported by a grant from the Spanish Ministry of Science and Technology (SEC2001-3821-C05-C02) with the collaboration of the European Fund for the Development of Regions. **Correspondence** should be sent to: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicologia. Carretera Valls s/n. 43007 Tarragona (Spain). E-mail: perejoan.ferrando@urv.net

stimulus is maximal when the stimulus is near the individual's psychophysical threshold (e.g. Guilford, 1954).

Some studies have used a classical test theory framework to address the DD hypothesis, by relating the respondent's (relative) total test score to the item's endorsement value -the percentage of respondents who endorsed the item (e.g. Kuncel, 1977). However, the hypothesis is better framed within a Thurstonian-type model in which the items and respondents can be represented by ordered points on the same underlying continuum of the trait of interest. This is because the item and the respondent locations in these models are on the same scale, and this provides a clear definition of the person-item distance. Conventional Item Response Theory (IRT) models such as the two-parameter or the one-parameter model are of this type; in particular, previous IRT-based research on this topic used the one-parameter logistic (Rasch) model (Kuncel, 1977; Kuncel and Fiske, 1974; Tyler, 1968).

Most of the previous research on the DD hypothesis used the stability over time as a correlate, and measured the instability by the proportion of responses which had changed between the first and the second administration**.** The results consistently suggest that the item response becomes more unstable under repetition as the person-item distance decreases (Ferrando, Lorenzo and Molina, 2001; Kuncel, 1977; Tyler, 1968).

Stability over time, however, is not the only dependent variable possible. Kuncel (1977) and Nowakowska (1983) hypothesized that longer response times are required when the person-item distance is small, and so the difficulty in responding is high. Again, this hypothesis is derived by analogy with the well-known psychophysical result that the response time increases as the stimulus approaches the individual's psychophysical threshold (e.g. Vickers, 1980).

In spite of its potential interest, the response time does not seem to have been used as a correlate for the distance-difficulty hypothesis, and this is the aim of the present study. At the item level, Hanley (1962) and Rogers (1973) found that the average response time tended to be smaller for items with extreme locations. This result, however, is not directly related to the present hypothesis. The DD hypothesis is relative, and states that the response time will tend to increase the nearer the respondent and the item are together, regardless of whether the respondents or items are located in extreme or non-extreme positions. The result, however, can be explained by the present hypothesis in the typical case that the distribution of the respondents' locations is normal or near normal. If this is so, most of

person-item distances will be large for an extreme item, and so the average response time is expected to be short.

At the intra-individual level, Kuncel (1977) reported a positive trend between nearness and response time for each respondent across a set of items. This approach, however, is problematic when the items are not equivalent. For example, a well-known empirical result is that the main determinant by far of the item response time is the length of the item stem (Dunn, Lushene and O'Neil, 1972; Rogers, 1974). It is unlikely that distance will have a clear effect across a set of items with different lengths, complexities, discriminatory powers, etc.

**Item Response Model, Distance Measure, and Design Issues**

The present study is intended for personality items that measure a continuous, dimensional personality trait, and in which the relation between the trait level and the probability of item endorsement is a dominance relation (see Coombs, 1964). Previous evidence suggests that personality measures of this type are well fitted by the two-parameter (2PM) or the one-parameter (1PM) IRT models (e.g. Ferrando, 1994; Finch and West, 1997; Reise and Waller, 1990; Waller, Tellegen, McDonald and Lykken, 1996). The 2PM and the 1PM are the models considered in this study

The person-distance measure habitually used in previous studies (Tyler, 1968, Kuncel, 1977) can be expressed in terms of the IRT model as:

$$\hat{\delta}^{(1)}_{ij=}\sqrt{(\hat{\theta}_i-\hat{b}_j)^2} \qquad (1)$$

where $\hat{b}_j$ is the estimated item location (threshold) and $\hat{\theta}_i$ is the estimated individual trait level. An alternative measure derived from the 2PM can be defined as:

$$\hat{\delta}^{(2)}_{ij=}\sqrt{\hat{a}_j^2(\hat{\theta}_i-\hat{b}_j)^2} \qquad (2)$$

where $\hat{a}_j$ is the estimated item discrimination (slope). In well designed items $\hat{a}_j > 0$ so the two measures can be related as: $\hat{\delta}^{(2)}_{ij=}\hat{a}_j\hat{\delta}^{(1)}_{ij}$ . Therefore, the second measure is a weighted version of the first. The role of the slope as a weight can be interpreted as follows. The item location can be

considered as the transition point at which respondents stop responding 'No' and start responding 'Yes'. Difficulty of responding is maximal when the individual trait value is the same as the item location, and in this case the probability of endorsing the item in the 2PM is 0.5. The slope controls the abruptness of the transition from the tendency to respond 'No' to the tendency to respond 'Yes': That is to say, the steeper the slope is, the more abrupt the transition, and the difficulty in responding is lower.

In the present study, the time-distance relations are studied at the intra-item level, i.e. over respondents within each of the items. This means that, for each analysis, the item characteristics are constant for all of the respondents. Note in particular that the threshold and slope item values are constant, which means that in this case the results will be the same with either of the distance measures discussed above. The measure of relation used in the analyses is the product-moment correlation ($r$) between the distances and the logs of the response times. Previous studies suggest that response times are usually well fitted by the lognormal distribution (Thissen, 1983; van der Linden and van Krimpen-Stoop, 2003). The logarithmic transformation normalizes the distributions of the response times and makes them more appropriate for the correlational analysis.

The within-item procedure used in this study is expected to remove the influence of the features that contribute most to the time taken to respond to the item, mainly item stem length and item complexity, because all these features are constant. Even so, the extent to which the relations will be strong can still not be predicted. If the DD-hypothesis is correct, then the person-item distance is one of the factors that contributes to the total item response time. However, there are likely to be some other individual-differences factors that consistently contribute to the response time, such as reading speeds, differences in understanding the item, differences in pressing the button, etc. Overall, the results of the present study attempt to give an idea of whether time-distance relationships exist or not (as predicted from the DD hypothesis). And, if they do, what the magnitude of these relations is.

Inferring the strength of time-distance relationships on the basis of a correlation coefficient presents two standard problems. First, the magnitude of $r$ depends on the homogeneity of the sample, so an attenuated estimate might be obtained if it is based on a convenience, possibly homogeneous, sample such as in the present study. Second, the distance measure is obtained from the item and individual parameter estimates, both of which are subject to measurement error. Therefore, the relationship between the estimated distances and the response times is not the same as the

relationship that would be obtained between the 'true' distances (i.e. those obtained from the item and individual parameter values) and the response times. From standard psychometric theory, what is expected is that the first relation be attenuated with respect to the second because of the measurement error.

This study also provides a disattenuated estimate of the correlation between the distances and the logs of the response times. This estimate would give us some idea of the extent of the relation between time and distance if there were no measurement error in the distance values. The correction for the attenuation procedure is explained in the Appendix.

## METHOD

**Measures and Participants**. The study used two short scales (Neuroticism, 11 items, and Extraversion, 11 items) selected from the Eysenck Personality Questionnaire Revised (EPQ-R, Eysenck, Eysenck and Barrett, 1985). Respondents were 286 Psychology and Social Sciences undergraduates at a university in Spain. The mean age was 19, and approximately 80% were female. They were asked to participate voluntarily in a research study and, at the end of the administration, they were provided with information about the test results.

**Procedures.** The items were administered by computer, and the general design of the computerized test followed the guidelines given by Kyllonen (1991). First the instructions were given. Then 25 items were presented: the first three were practice items and were used so that the students could warm up and familiarize themselves with the task. The next 22 were a mixture of the items of the two scales. The response system consisted of two central response buttons (YES/NO), and one button on the right for obtaining the next item. At the end of the administration a window appeared to thank the respondents for their cooperation.

The computer recorded the time in milliseconds from the appearance of the item on the screen to the moment the corresponding response button was pressed. Next the logarithmic transformation was applied to the direct response times.

# RESULTS

### Item calibration and scoring

The item responses were fitted using both the 2PM and the 1PM. The item parameters were estimated with the BILOG program (Mislevy and Bock, 1990) using the Marginal Maximum a Posteriori procedure. The scale metric was obtained by setting the trait distribution to zero mean and unit variance. The goodness of fit model-data was assessed by using the item factor analysis parameterization of the 2PM and the 1PM as implemented in the LISREL program (Jöreskog and Sörbom, 1996). This approach provides a detailed assessment of the goodness-of- fit results. Furthermore, it is possible to compare the fit of the 2PM and the 1PM by means of a hierarchical test, by restricting the slopes of the 2PM to be equal. The goodness of fit results are summarized in Table 1.

**Table 1. Item Calibration: Goodness of Fit Results.**

Neuroticism Scale

| Model | ε-RMSEA | 90% C.I. | γ-RDR | NNFI |
|---|---|---|---|---|
| 2PM | 0.068 | (0.051;0.086) | | 0.97 |
| 1PM | 0.110 | (0.096;0.125) | 0.220 | 0.92 |

Extraversión Scale

| Model | ε-RMSEA | 90% C.I. | γ-RDR | NNFI |
|---|---|---|---|---|
| 2PM | 0.054 | (0.034;0.073) | | 0.97 |
| 1PM | 0.116 | (0.101;0.131) | 0.251 | 0.87 |

Note: ε-RMSEA, point estimate of the root mean squared error of approximation; 90% C.I. 90% Confidence Interval for the RMSEA, γ-RDR, root deterioration per restriction; NNFI, non-normed fit index.

The results in Table 1 are quite clear. According to the point and interval estimates of the RMSEA as well as the values of the non-normed fit index, the 2PM has an acceptable fit in both scales. The fit of the 1PM, however, must be judged to be unacceptable on both scales (see Hu and

Bentler, 1999 for cut-off criteria). As for the nested comparison between both models, the root deterioration per restriction suggests that, in both scales, the fit worsens dramatically when the restriction of equal slopes is imposed. Browne and Du Toit (1992) suggested that an RDR value below 0.05 is needed if it is to be considered that the restrictions do not significantly worsen the fit. Overall, it appears that the model that should be used is the 2PM.

The respondent parameters (i.e. individual trait levels) were next estimated based on the 2PM item estimates using the robust biweight procedure, and then rescaled to have zero mean and unit variance, and so to be on the same scale as the item locations (see Mislevy and Bock, 1990). Next, the results obtained in this scoring stage were used to compute the reliability estimate for the distances (see the Appendix). The estimated reliabilities were: 0.80 (Neuroticism scale) and 0.64 (Extraversion scale).

**Analysis of the time-distance relation.**

Table 1 presents the time-distance relations separately for both scales. All the correlations are negative, as expected from the theory. To assess the significance of the results, first the omnibus null hypothesis that the vector of correlations is zero in the population was tested by means of Steiger's (1980) quadratic form asymptotic chi-square statistic. The results were: $\chi^2$=31.31, *df*=11 *p*=0.001(Neuroticism scale), and $\chi^2$=51.68, *df*=11, *p*=0.00001 (Extraversion scale). In both cases, the null hypothesis was rejected. The corresponding effect sizes ($\chi^2/N$) were 0.11 and 0.18.

The columns on the right of the correlations show the estimated 90% confidence intervals, and the columns on the far right-hand side show the disattenuated correlation estimates. The confidence intervals were obtained by means of Bootstrap resampling based on 2000 replications. For 15 of the 22 items, both ends of the confidence interval are negative. This is equivalent to considering that the corresponding correlations differ significantly from zero according to a conventional one-tailed 5% significance test. It should be taken into account however that the confidence intervals are only orientative. The simple bootstrap design used here treats the distances as observed values, although they are in fact obtained from fallible estimates which also have sampling variability. So, the confidence intervals are probably somewhat wider than those reported in table 2.

As for differential effects, the expected relations appear to be somewhat stronger for the Extraversion items, as the omnibus test and the disattenuated correlations also suggests. To assess this point in more detail,

the correlation between the average response time and the average distance over items was computed in both of the scales. The resulting correlations between averages were *r*=-0.20 (Neuroticism) and *r*=-0.24 (Extraversion).

**Table 2. Product-Moment Correlations Between the Person-Item Distances and the Logs of the Response Times for both the Neuroticism and the Extraversion Scales**

| Neuroticism | | | | Extraversion | | | |
|---|---|---|---|---|---|---|---|
| Item | r | 90% C.I. | r-dis | Item | R | 90% C.I. | r-dis |
| N1 | -0.06 | (-0.10;0.09) | -0.07 | E1 | -0.27 | (-0-36;-0.17) | -0.34 |
| N2 | -0.22 | (-0.30;-0.12) | -0.25 | E2 | -0.15 | (-0.25;-0.05) | -0.19 |
| N3 | -0.10 | (-0.19;0.00) | -0.11 | E3 | -0.22 | (-0.31;-0.12) | -0.27 |
| N4 | -0.11 | (-0.20;-0.01) | -0.12 | E4 | -0.14 | (-0.24;-0.04) | -0.17 |
| N5 | -0.09 | (-0.17;0.01) | -0.10 | E5 | -0.23 | (-0.32;-0.13) | -0.29 |
| N6 | -0.11 | (-0.20;-0.01) | -0.12 | E6 | -0.05 | (-0.15;0.05) | -0.06 |
| N7 | -0.20 | (-0.28;-0.10) | -0.22 | E7 | -0.25 | (-0.34;-0.15) | -0.31 |
| N8 | -0.16 | (-0.25;-0.06) | -0.18 | E8 | -0.13 | (-0.23;-0.03) | -0.16 |
| N9 | -0.13 | (-0.21;-0.03) | -0.15 | E9 | -0.03 | (-0.12;0.07) | -0.04 |
| N10 | -0.10 | (-0.19;0.00) | -0.11 | E10 | -0.20 | (-0.29;-0.10) | -0.25 |
| N11 | -0.19 | (-0.28;-0.09) | -0.21 | E11 | -0.01 | (-0.10;0.09) | -0.01 |

Note: r, product-moment correlation; 90% C.I., Bootstrap 90% Confidence Interval for r; r-dis, disattenuated correlation.

## DISCUSSION

Overall, the results tend to support the hypothesis that the response time increases the closer the item and respondent locations are together. However, the relations obtained are quite weak. The Pearson correlation coefficient is itself an effect size estimate, and according to Cohen's (1977) conventional definition, all the effect sizes obtained in this study are small. As discussed above, the correlation values obtained in this study could be attenuated. First, the sample is a convenience sample of Social Science undergraduates, and might be considered to be homogeneous as far as the Neuroticism and Extraversion levels is concerned. Second, the distances are computed from fallible item and respondent estimates. However, it is clear that even if we correct for unreliability, the disattenuated estimates are still small. An 'ideal' study based on a large random sample and a large set of items (so as to provide a more accurate estimate of each individual level) could give a more accurate idea of the strength of the relation. Also, because the strength of the relation may depend on the type of trait that is measured, the study should be extended to a larger number of traits. Given

the weak results obtained in this study, however, the selection of traits and measures should be theoretically guided, and only those traits and instruments for which a clear relation is expected should be investigated. This is left for future research.

We shall now discuss the potential usefulness of the results obtained in this study. First, we note that in a computerized administration of a questionnaire the item response time can be obtained at no additional cost (in any sense) as an auxiliary source of information that complements the standard item response. In contrast, additional measures of information such as stability require repeated administration of the questionnaire. Now, if the response time is consistently related to certain item and respondent characteristics, as the present results suggest, then the additional information provided by this variable could be used to obtain more accurate estimates of the item and respondent parameters. This is the role of response time in some latency models previously proposed in the literature (Thissen, 1983; van der Linden and van Krimpen-Stoop, 2003) for the ability domain. It would be interesting to develop a similar model for personality items that incorporates the information provided by the response time for each of the items.

## RESUMEN

**Tiempo de respuesta y distancia entre la persona y el ítem: Un estudio empírico en personalidad.** El presente estudio evalúa la hipótesis de que el tiempo de respuesta a un ítem aumenta a medida que las posiciones del sujeto y del ítem en el continuo que se mide se van haciendo más próximas. Esta hipótesis se ha propuesto repetidas veces en la literatura pero no parece haber sido nunca evaluada de forma rigurosa. En este estudio se administró una versión computerizada de un cuestionario de personalidad (2 escalas con 11 ítems cada una) a una muestra de 286 sujetos. Los ítems se calibraron mediante un modelo de TRI, y se obtuvo una medida de distancia derivada de dicho modelo. A continuación se calcularon dentro de cada ítem las correlaciones producto-momento entre las distancias y los logaritmos de las latencias. En ambas escalas todas las correlaciones fueron negativas, tal como plantea la teoría. Sin embargo, los valores de correlación fueron bastante bajos. Se incluye una discusión de la potencial utilidad que tienen los resultados obtenidos para la medición en personalidad.

# REFERENCES


Browne, M.W. and Du Toit, S. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27*, 269-300.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Coombs, C.H. (1964). *A theory of data*. New York: Wiley.

Dunn, T.G.; Lushene, R.E., and O'Neil, H.F. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology, 39*, 381-387.

Eisenberg, P. and Wesman, A.G. (1941). Consistency in response and logical interpretation of psychoneurotic inventory items. *Journal of Educational Psychology, 32*, 321-338.

Eysenck, S.B.G., Eysenck, H.J., and Barrett, P.T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences. 6,* 21-29.

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, U. and Molina, G. (2001). An Item Response Theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25*, 3-17.

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

Guilford, J.P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

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

Hu, L. and Bentler, P.M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jöreskog, K.G. and Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software.

Kuncel, R.B. (1977). The subject-item interaction in itemmetric research. *Educational and Psychological Measurement, 37,* 665-678.

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

Kyllonen, P.C. (1991). Principles for creating a computerized test battery. *Intelligence, 15*, 1-15.

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

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

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

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

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

Rogers, T.B. (1974). An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica, 38*, 205-213.

Steiger, J.H. (1980). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research, 15,* 335-352.

Thissen, D. (1983). Timed testing: An approach using Item Response Theory. In D.J.Weiss (Ed.), *New Horizons in Testing* (pp. 179-203). New York: Academic Press.

Tyler, T.A. (1968). *Response stability, person-item distance, and homogeneity*. Unpublished doctoral dissertation, University of Chicago.

Van der Linden, W.J. and van Krimpen-Stoop, E.M.L.A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika,* 251-265.

Vickers, D. (1980). Discrimination. In A.T. Welford (Ed.) *Reaction Times* (pp. 25-72). New York: Academic Press.

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

# APPENDIX

## Correction for attenuation

First, a conventional measure of reliability for the distance is defined as (e.g. Lord, 1980)

$$\rho_{\delta\delta} = 1 - \frac{E(Var(\hat{\delta} \mid \delta))}{Var(\delta)} \qquad (1)$$

where the expectation of the conditional error variance is taken over the distribution of the distances. Next, we assume that the item parameters are fixed and now values, and then use a first-order Taylor approximation (i.e. the delta method). We obtain

$$\begin{aligned} Var(\delta) &\cong Var(\theta) \\ Var(\hat{\delta} \mid \delta) &\cong Var(\hat{\theta} \mid \theta) \end{aligned} \qquad (2).$$

So:

$$\rho_{\delta\delta} \cong 1 - \frac{E(Var(\hat{\theta} \mid \theta))}{Var(\theta)} \qquad (3).$$

Expression (3) can be computed as a by-product of the parameter estimation procedure. The expectation of the conditional variance is obtained from the average information function, whereas $Var(\theta)$ is usually fixed to unity. Finally, the disattenuated correlation is obtained by dividing the time-distance product-moment correlation by the square root of the reliability estimate (3).

(Manuscript received: 30 June 2004; accepted: 24 February 2005)