Psicothema

Psicothema 2021, Vol. 33, No. 2, 259-267
doi: 10.7334/psicothema2020.346

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Reliability and External Validity of Personality Test Scores: The Role of Person and Item Error

Pere J. Ferrando, and David Navarro-González
Universitat Rovira i Virgili

Abstract

Background: This article explores the suitability of a proposed "Dual" model, in which both people and items are sources of measurement error, by assessing how the test scores are expected to behave in terms of marginal reliability and external validity when the model holds. **Method:** Analytical derivations are produced for predicting: (a) the impact of person and item errors in the amount of marginal reliability and external validity, as well as the occurrence of "ceiling" effects; (b) the changes in test reliability across groups with different average amounts of person error, and (c) the phenomenon of differential predictability. Two empirical studies are also used both as an illustration and as a check of the predicted results. **Results:** Results show that the model-based predictions agree with existing evidence as well as with basic principles in classical test theory. However, the additional inclusion of individuals as a source of error leads to new explanations and predictions. **Conclusions:** The proposal and results provide new sources of information in personality assessment as well as of evidence of model suitability. They also help to explain some disappointing recurrent results.

Keywords: Personality measurement, person fluctuation, item discrimination, marginal reliability, external validity, differential predictability.

Resumen

Fiabilidad y Validez Externa de las Puntuaciones en los Tests de Personalidad: El Papel del Error en las Personas y en los Ítems. **Antecedentes:** se explora la adecuación de un modelo "Dual" en el que ítems y personas son fuente de error de medida, evaluando como se espera que se comporten las puntuaciones en un test en términos de fiabilidad y validez cuando el modelo se cumple. **Método:** se derivan analíticamente predicciones respecto a: (a) el impacto del error en personas y en ítems en las estimaciones de fiabilidad y validez externa, así como en efectos techo esperados, (b) cambios en la fiabilidad marginal en grupos con diferente magnitud media de error individual, y (c) el fenómeno de la predictibilidad diferencial. Se incluyen dos estudios empíricos a efectos de ilustración y verificación empírica. **Resultados:** las predicciones concuerdan con la evidencia acumulada y con los principios de la teoría clásica del test. Sin embargo, la inclusión del parámetro de error individual permite llegar a nuevas explicaciones y predicciones. **Conclusiones:** la propuesta y resultados proporcionan nuevas fuentes de información en la medida de la personalidad, así como evidencia de la adecuación del modelo. También explican algunos resultados decepcionantes y recurrentes.

Palabras clave: medición de la personalidad, fluctuación individual, discriminación del ítem, fiabilidad marginal, validez externa, predictibilidad diferencial.

When compared to ability scores, personality scores usually show weaker psychometric properties in several aspects (Fiske, 1963; Hofstee et al., 1998; Morgeson et al., 2007). In particular, they (a) typically have lower conditional and marginal reliabilities (e.g. Hofstee et al., 1998), and, (b) their validity relations with other relevant external variables (criteria or other test scores; e.g. Muñiz & Fonseca-Pedrero, 2019; Siresi & Padilla, 2014) are generally disappointingly weak (Morgeson et al., 2007; Paunonen & Jackson, 1985). We shall denote here this source of validity evidence as "external validity".

Some authors consider that vague conceptualizations, poor designs, poorly developed items, and poor scoring schemas are, among other deficiencies, at the root of the problem (Hofstee et al., 1998; Fiske, 1963; Muñiz & Fonseca-Pedrero, 2019; Morgeson et al., 2007; Paunonen & Jackson, 1985). However, evidence suggests that these improvements would increase accuracy and external validity only to a certain extent, and that there seems to be a sort of 'ceiling' that cannot be surpassed no matter how much the items are improved, how well the test is designed, or how "optimal" the chosen scoring schema is (Hofstee et al., 1998; Loevinger, 1957; Taylor, 1977).

Most item response theory (IRT) models that are used in personality measurement consider the items as the sole source of error in the test score. However, since the 1940s this view has been challenged by several authors (Coombs, 1948; Ferrando, 2019; Fiske, 1963; Lumsden, 1978, 1980; Mosier, 1942). In particular, Lumsden (1978) proposed an alternative view in which items were perfectly reliable and the sole source of measurement error were the respondents. This view, however, leads to predictions that are hard to match with the evidence collected so far. If respondents were the sole source of measurement error, conventional analysis

Received: September 15, 2020 • Accepted: November 25, 2020
Corresponding author: Pere J. Ferrando
Facultad de Psicología
Universitat Rovira i Virgili
43007 Tarragona (Spain)
e-mail: perejoan.ferrando@urv.cat

of personality items would invariably result in equal discriminating power values for all the test items, which is not the case (Ferrando, 2013). Furthermore, the reliability of the test scores would not be expected to increase if the test were lengthened. However, the predictions given by the Spearman-Brown prophecy seem to work well in personality applications.

In further articles, Lumsden (1980) modified his viewpoint and adopted a position that can be summarized in two points. First, it is no longer assumed that all the items in a test are error-free, but only that they all have the same amount of error. Second, this assumption is not expected to occur "per se". Rather, a careful process of item selection is generally required. With these modifications, Lumsden's 1980 model becomes a restricted version of a more general model in which both items and persons are characterized by different amounts of error. This general model, which is that considered here, is regarded as the most realistic for fitting personality measures by the authors referred to above.

Ferrando (2013, 2019) provided what appears to be the first workable framework for modeling responses with errors in both persons and items. The proposed general framework uses IRT modeling based on a Thurstonian response mechanism, and includes specific person and item parameters for modeling the amounts of measurement error. So far, efforts have mainly focused on developing the model. However, there are many aspects that still require further research, and this article focuses on some of them.

The main purpose of this article is to explore the predictions that can be made by the modeling above in terms of the two general properties of test scores: (marginal) reliability and external validity. In more detail, we wish to assess how the scores are expected to behave in terms of reliability and validity as a function of the model parameters and distributional assumptions. Simple-sum test scores are, by far, the most commonly used in psychometrics (e.g. Hontangas et al., 2016; Muñiz & Fonseca-Pedrero, 2019). So, assessing how they are expected to behave under the modeling proposed seems to be of practical interest. More specifically, the present study is relevant for four main reasons. First, it assesses whether the model-based predictions agree with standard results in psychometrics and existing evidence. Second, it provides a meaningful explanation for the 'ceiling' limitations that personality test scores appear to have in terms of reliability and validity. Third, it explores the role that the new model parameters have in predicting how the test scores function. Finally, it assesses model appropriateness on the basis of predicted-observed outcomes.

The comprehensive framework proposed by Ferrando (2019) included models (or sub-models) intended for continuous, graded, and binary responses. We have chosen here as a starting point the simplest model intended for continuous responses, which has been named *Dual Thurstonian Continuous Response Model* (DTCRM; Ferrando, 2019). On the one hand, both, the DTCRM and the classical test theory (CTT) concepts of reliability and validity are based on direct analyses of item and test scores, and assume that the relations between these scores and the trait levels are linear. So, the relations obtained here are relatively direct, clear and interpretable, and can be expressed in closed form. On the other hand, response formats that approach continuity (e.g. line segments or visual analogue scales) are relatively common in personality, and are generally well fitted by a continuous model. Furthermore, more conventional graded formats in 5 or more points can also be expected to be reasonably well approximated by this model, because the item distributions in the personality domain are generally unimodal and not too extreme (see Ferrando, 2013, for a detailed discussion). Indeed, it would be of interest to extend the present results to the categorical-response models in future developments.

A Review of the DTCRM

Consider a test made up of $j= 1 \ldots . n$ items with an approximately continuous format (a line segment, a visual analogue scale, or a graded response scale with the distributional properties discussed above. The scale is assumed to be the same for all the test items. We shall also assume that the format endpoints are labeled "totally disagree" and "totally agree", so that a greater item score means a greater degree of agreement or endorsement. The n test items aim to measure a trait θ, which is assumed to have zero mean and unit variance.

According to the DTCRM, at the moment of encountering item j, respondent i has a momentary perceived trait level T_i. At this moment, item j has a momentary location b_j. These momentary values are modeled as

$$T_i = \theta_i + \omega_i : b_j = \beta_j + \varepsilon_j \quad (1)$$

The distribution of T_i over the test items is normal, with mean θ_i and variance σ^2_i. The θ_i, value is the central trait value or the person location, the single value that best summarizes the standing of individual i in the θ trait (Mosier, 1942). The amount of fluctuation of the momentary perceived values around θ_i reflects the error of respondent i, and is summarized by variance σ^2_i. Although we shall work here directly with the σ^2_i parameter, from a substantive point of view it is useful to consider its reciprocal ($1/ \sigma^2_i$) as a measure of person reliability (e.g. Lumsden, 1978) or, in Thurstone's (1927) terminology, of person discrimination. The magnitude of ($1/ \sigma^2_i$) is thought to reflect mainly the relevance and degree of clarity and strength with which the trait is internally organized in the individual (e.g. LaHuis et al., 2017). So, individuals for which θ is highly relevant and well organized are expected to respond with high sensitivity and discrimination to the different trait manifestations sampled by the test items. The fluctuation error around the central value would then be low, and the person reliability high.

The distribution of the momentary item location b_j across respondents is also normal, with mean β_j, and variance $\sigma^2_{\varepsilon j.}$ The dispersion of the b_j values around the central item location β_j, models the item error, and, again, closely corresponds to Thurstone's (1927) concept of discriminal dispersion. It is summarized by the variance $\sigma^2_{\varepsilon j}$, and, its reciprocal ($1/ \sigma^2_{\varepsilon j}$) would be a measure of item discriminating power. The different amounts of $\sigma^2_{\varepsilon j}$, are expected to reflect mainly (a) 'surface' item characteristics, particularly length and verbal complexity, and (b) 'itemmetric' characteristics, particularly ambiguity and trait indicativity (e.g. Ferrando & Demestre, 2008).

Let X_{ij} be the score of individual i in item j. The structural model for this score is

$$X_{ij} = \gamma + \lambda(T_i - b_j) \quad (2)$$

The γ parameter in (2) is the response scale midpoint, whereas λ is a scaling parameter, which has a positive value, and relates the item score scale to the latent scale of θ. So, γ and λ are simply intercept and scale parameters. The difference T_i-b_j is the momentary

person-item distance (see Ferrando, 2013) and determines both the direction and extremeness of the response. So, when $T_i > b_j$, the model-implied score is above the response midpoint (i.e. in the agreement direction). And, the greater the distance is in any direction (above or below the response midpoint), the greater the response extremeness is.

At the methodological level, the DTCRM is fully described in Ferrando (2013, 2019), and no further details of this type will be provided here. It is implemented in InDisc, an R package (Ferrando & Navarro-González, 2020; Navarro-González & Ferrando, 2020) that can be obtained from https://cran.r-project.org/package=InDisc. A complete user's guide can be found at http://psico.fcep.urv.cat/utilitats/InDisc.

Results at the Single-Item Score level

The correlation between the scores in item j and θ according to the DTCRM is

$$\rho_{Xj\theta}=\frac{1}{\sqrt{1+E\left(\sigma_i^2\right)+\sigma_{\varepsilon j}^2}} \quad (3)$$

From standard CTT principles, index (3) can be interpreted as an item reliability index: i.e. the correlation between the observed item scores and the 'true' trait levels (Mellenbergh, 1994). So, again following Mellenbergh (1994), the square of (3) can be interpreted as the reliability coefficient of the item's j scores:

$$\rho^2 Xj\theta = \rho XjXj = \frac{1}{1+E\left(\sigma_i^2\right)+\sigma_{\varepsilon j}^2} \quad (4)$$

The most interesting results from (4) concern the role of person and item error in the amount of single-item-score reliability that can be attained. This amount depends on both the average of the person fluctuations in this group, and the amount of dispersion of this specific item. So, even if a "perfect" item with zero dispersion could be designed, the reliability of the item's scores would still be below unity due to an irreducible source of error that would reflect the inherent fluctuations of the respondents. More generally, the unreliability of the item scores is expected to reflect the combined effect of both sources of error: persons and item. For the values usually obtained so far with personality items (Ferrando, 2013, 2019), figure 1 shows the expected item score reliability as a function of both sources.

Figure 2 is the diagram of equation (4) for different values of item dispersion, so it can be viewed as "slices" of figure 1 at fixed values on the x axis. The resulting curves are rectangular hyperbolas with different degrees of curvature. The upper curve corresponds to a 'perfect item', and shows how in this case item score reliability strongly decreases as person fluctuation increases. The lowest curve corresponds to a "noisy" item and, in this case, the impact of person fluctuation is much smaller, as the item scores are highly unreliable in all cases.

Although not operatively formalized, the relations discussed so far have already been considered in the literature. Taylor (1977) suggested that the reliability of personality items were partly intrinsic and could not be explained by item ambiguity or other item characteristics. And Lumsden (1978) also considered that the amount of individual fluctuation might well be different in different groups. Now, in our model, the amount of person error is operationalized as the average of the person fluctuations. The reliability "ceiling" imposed by this source is that given in (4) when $\sigma_{\varepsilon j}^2=0$, and two interesting conjectures are that this ceiling

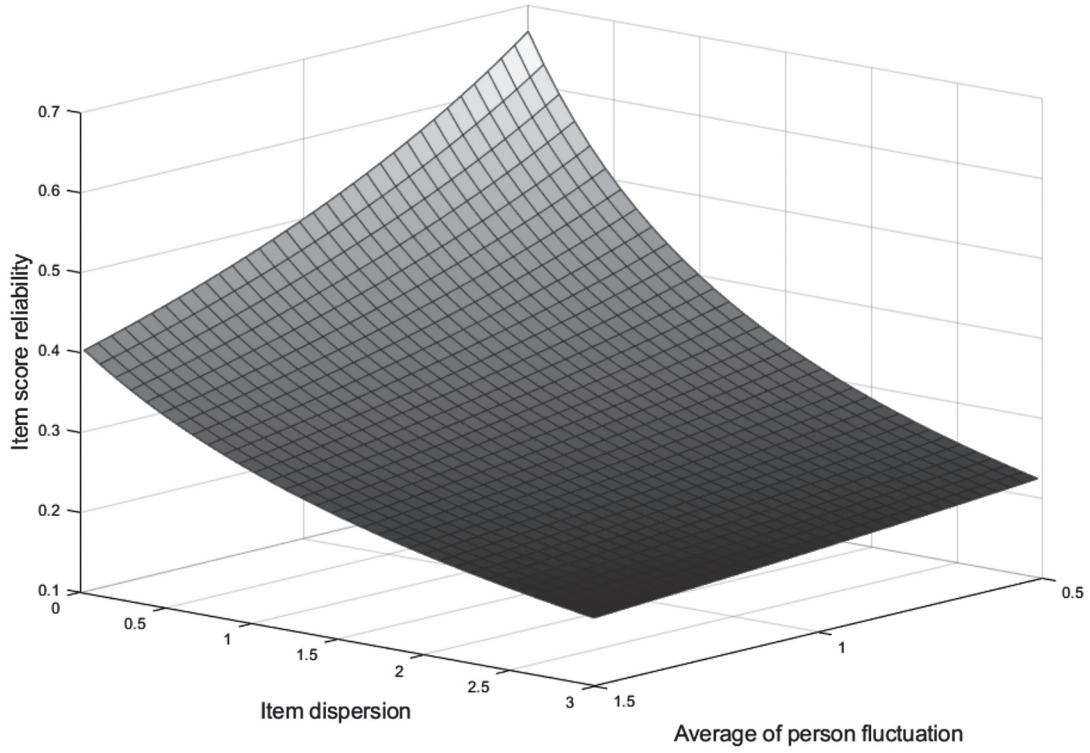

Figure 1. Expected item score reliability as a function of item dispersion and average person fluctuation

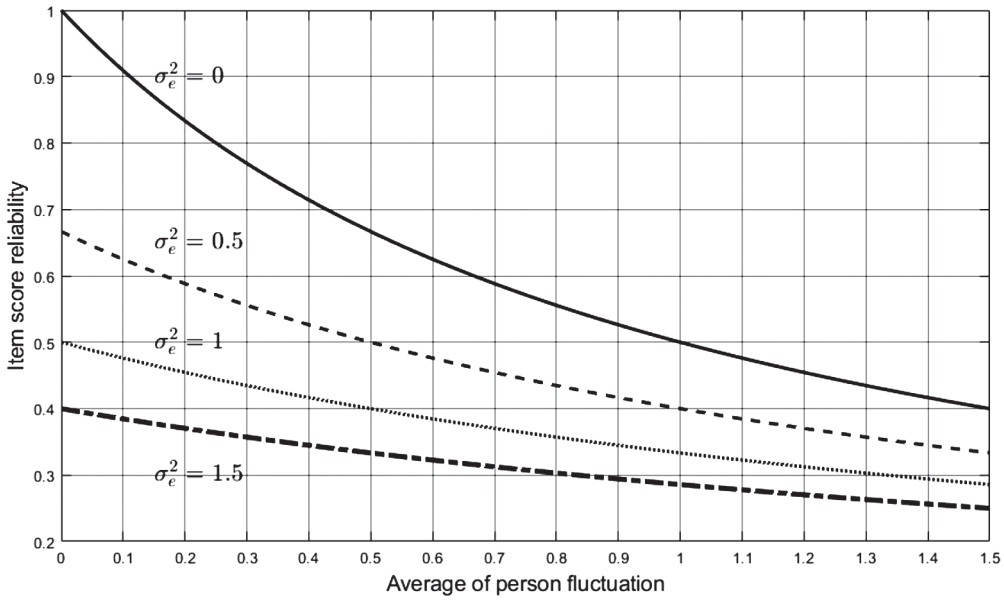

Figure 2. Contour plot representing the expected item score reliability as a function of the average of person fluctuation for four different amounts of item dispersion

might well depend on (a) the type of trait that is measured and (b) the group of individuals that is assessed.

The correlation between the scores on a pair of items *j* and *k* is now found to be:

$$\rho XjXk = \frac{1}{\sqrt{1+E\left(\sigma_i^2\right)+\sigma_{\varepsilon j}^2}\sqrt{1+E\left(\sigma_i^2\right)+\sigma_{\varepsilon k}^2}} \tag{5}$$

Note further that in the case of "parallel' items with the same amount of dispersion ($\sigma_{\varepsilon j}^2 = \sigma_{\varepsilon k}^2$), the model-implied correlation (5) becomes the item reliability in (4). This result is consistent with the CTT definition of reliability as the correlation between two equivalent measures (e.g. Muñiz, 2000).

As in (4), the average person fluctuation places a ceiling on the maximum inter-item correlation that can be attained (i.e. if both items were "perfect"), an idea that has already been suggested: Loevinger (1957) considered that there is a particular upper limit of inter-correlations among trait manifestations (item scores in this case), and named this upper limit 'characteristic intercorrelation'.

The Reliability of Test Scores as a Function of the DTCRM Parameters

The two most standard definitions of reliability (Lord & Novick, 1968; Muñiz, 2000) are based on (a) the squared correlation between true and observed scores, or (b) the ratio of true variance to observed variance. It can be shown (further details can be obtained from the authors) that both definitions led here to the same result: The reliability of the test score as implied by the DTCRM is found to be:

$$\rho_{XX} = \frac{1}{1+\dfrac{E_i\left(\sigma_i^2\right)}{n}+\dfrac{E_j\left(\sigma_{\varepsilon j}^2\right)}{n}} \tag{6}$$

And clearly allows its determinants to be assessed. They are: (a) test length, (b) average (over respondents) of the person fluctuations, and (c) average (over items) of the item dispersions. The role of (b) and (c) has already been discussed in (4). As for (a), equation (8) predicts that (other factors constant) reliability increases with test length, which is consistent with conventional CTT wisdom.

We shall now study in more depth the extent to which the DTCRM-based predictions are consistent with basic CTT results. If we consider again the case of a test made of parallel items, equation (6) can be written as

$$\rho_{XX} = \frac{1}{1+\dfrac{E_i\left(\sigma_i^2\right)}{n}+\dfrac{E_j\left(\sigma_\varepsilon^2\right)}{n}} = \frac{n\rho_{jk}}{1+(n-1)\rho_{jk}}$$
$$= \left(\frac{n}{n-1}\right)\frac{Var(X)-\sum Var(X_j)}{Var(X)} \tag{7}$$

where ϱ_{jk} is the common correlation between any pair of items. The two expressions on the right-hand side of (7) are well-known expressions for Cronbach's (1951) alpha reliability coefficient. Further, if we consider a test made of *n* parallel items that conform to the DTCRM, and if this test is lengthened *m* times by adding equivalent items, the predicted reliability of the lengthened test scores is:

$$\rho_{XX(m)} = \frac{m\rho_{XX}}{1+(m-1)\rho_{XX}} \tag{8}$$

The well-known Spearman-Brown prophecy. So, if the DTCRM holds, the expected results as far as test score reliability is concerned are fully consistent with basic CTT results. The most interesting results, however, are those concerned with the role that the amount of person fluctuation plays in the expected reliability of the test scores. Consider a personality test characterized by the

number of items with their central locations and dispersions. As in Lumsden (1978), we consider that these characteristics depend only on the test, so they are assumed to remain invariant if the test is administered in different groups. Consider next the assumption that the average person fluctuation varies across the different groups to which the test can be administered. In this case, the expected reliability of the test scores in each group is a sole function of the average group fluctuation, and can be predicted according to (6). Among other things, this result leads to an internal procedure for checking model appropriateness: Individuals can be sorted by their σ^2_i estimates and assigned to intervals on the σ^2_i continuum on the basis of their estimated values. Next, the expected reliability of the test scores in each interval can be obtained from (6) by using the mean of the estimated fluctuation values in the interval. At the same time, the empirical reliability of the scores in each interval can be directly obtained by using standard procedures (e.g. alpha estimates if the item dispersions are not too different). Agreement between the observed-expected reliabilities in each interval would then provide support for the appropriateness of the model.

The external validity of test scores as a function of the DTCRM parameters

Consider now an external variable Y that is related to the central values θ (see Grucza & Goldberg, 2007; and Muñiz & Fonseca-Pedrero for potential outcomes that play the role of Y). We shall first define the correlation between Y and the central values θ ($\varrho_{\theta Y}$) as the theoretical validity coefficient, and the usual correlation between Y and the X test scores (ϱ_{XY}) as the empirical validity coefficient (Lord & Novick, 1968, sect. 12.1). As a function of the DTCRM parameters and distributional assumptions, the relation between both coefficients is readily found to be

$$\rho_{XY} = \frac{\rho_{\theta Y}}{\sqrt{1 + \frac{E_i\left(\sigma_i^2\right)}{n} + \frac{E_j\left(\sigma_{\varepsilon j}^2\right)}{n}}} \tag{9}$$

So, the empirical validity coefficient is an attenuated estimate of the theoretical validity coefficient, and the factors that determine the amount of attenuation are again (a) the average of the individual fluctuations, (b) the number of items, and (c) the average item dispersion. Now, if we combine (6) and (9), expression (9) can be written as:

$$\rho_{\theta Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX}}} \tag{10}$$

which is the standard CTT correction-for-attenuation formula when only the predictor (test score) is corrected for error (Lord & Novick, 1968; Muñiz, 2000). So, again, the validity predictions based on the DTCRM are consistent with the basic CTT results.

To find the new validity contributions allowed by the model, consider as before a personality test in which both item locations and item dispersions remain invariant. Then, according to (9) the amount of validity attenuation is a sole function of the average person fluctuation, and the empirical validities would become more and more attenuated as the person fluctuation of the group which is tested increases. Furthermore, consider now different groups (or sub-groups) of individuals with different average fluctuations. Because the expected empirical validity in each sub-group is a sole function of the average fluctuation, the "predictability" of each sub-group is itself predictable. To make this idea more specific, consider two subgroups A and B, and let $\varrho_{XY}^{(A)}$ be the empirical validity coefficient in group A. The model-expected empirical validity in group B is then given as:

$$\rho_{XY}^{(B)} = \rho_{XY}^{(A)} \sqrt{\frac{1 + \left[\frac{E_i\left(\sigma_i^2\right)^{(A)}}{n + E_j\left(\sigma_{\varepsilon j}^2\right)}\right]}{1 + \left[\frac{E_i\left(\sigma_i^2\right)^{(B)}}{n + E_j\left(\sigma_{\varepsilon j}^2\right)}\right]}} \tag{11}$$

Results (9) and (11) provide tools for (a) making tangible predictions about proposals or conjectures that have been made in the personality literature and in personnel selection, and (b) Checking model appropriateness. As for point (a) it has been proposed that some individuals (predictable individuals) have smaller errors of prediction with regards to specified criteria, while others are far less predictable and are characterized by large prediction errors; a proposal known as "differential validity" (e.g. Ghiselli, 1963). Some authors (e.g. Berdie, 1961) further noted that differential validity could, in turn, be predicted from individual differences in variability (i.e. fluctuation). However, no formal, model-based predictions based on this idea appear to have been proposed so far. In our modeling, however, σ_i^2 would be regarded as a predictability index, and expected differences in predictability could be estimated from equations (9) and (11).

As for point (b) above, the results in this section can be used to check the appropriateness of the model, similar to that described in the reliability section, by comparing the agreement between the observed and predicted (using equation 11) empirical validity coefficients in each of the intervals obtained.

Illustrative Examples

Example 1: Reliability

In this example we used the CTAC questionnaire, the Spanish acronym for "Anxiety Questionnaire for Blind People" (see Ferrando, 2019). The CTAC is a 35-item test that measures anxiety in situations related to visual deficit and which is used in the general adult population with severe visual impairment. The response format is 5-point Likert and, in the population for which the test is intended, the distributions of the item scores are generally unimodal and not extreme, which makes the DTCRM an appropriate model. This questionnaire was administered to a sample of 758 individuals with visual impairments. The dataset we are summarizing has already been analyzed with the DTCRM, and the details on calibration and model-data fit results can be found in Ferrando (2019). So, the only results provided here are those needed to illustrate the role of person fluctuation in the reliability of the CTAC scores. The estimated averages of the item dispersions and the person fluctuations were 1.10 and 1.06, respectively, and, with these results, the predicted marginal reliability of the CTAC scores in the entire sample using equation (6) was 0.942, which closely agrees with the empirical alpha estimate of 0.946.

Next, the schema explained above was used to predict changes in the marginal reliability as a function of average person fluctuation according to equation (6). Given the sample size, individuals were assigned to 7 intervals on the σ^2_i continuum with about 100 individuals per interval. In each interval, we computed (a) the expected reliability, obtained from (6) using the mean fluctuation, and (b) the empirical reliability using the alpha estimate. Results are in table 1 and depicted in figure 3.

It is clear that the reliability of the CTAC scores decreases as average fluctuation increases, as the model predicts. Furthermore, the observed-expected agreement is quite acceptable in the first five intervals. In the last two, however, the decrease in the empirical reliability is much more pronounced than the model predicts. This result suggests that the most extreme intervals contain not only the respondents with the largest amounts of fluctuation, but also the individuals who respond inconsistently for other reasons.

Example 2: Validity

In this example, a 30-item extraversion (E) scale which was administered to 338 undergraduate students was used. The item stems were taken from different Eysenck questionnaires, and the response format was 5-point Likert. Additionally, a short impulsivity scale of four items was also administered and taken as the external variable to be predicted from the E scores.

Table 1
Reliability estimates for the seven intervals on the σ^2_i continuum. Example 1

Empirical reliability	0.99	0.98	0.96	0.94	0.92	0.86	0.73
Predicted reliability	0.97	0.96	0.95	0.94	0.93	0.92	0.89
Average person fluctuation	0.05	0.27	0.58	0.88	1.25	1.70	2.99

Using InDisc, the σ^2_i estimates based on the E scale were computed for the 338 respondents. Next, respondents were sorted according to their estimates, and assigned to one of the five intervals with 67-68 participants in each. In each interval, we then computed (a) the predicted validity, obtained from (11) by using the mean fluctuation in each interval, and (b) the empirical validity as defined above. Results are in table 2 and depicted in figure 4.

As the model predicts, empirical validity decreases as average fluctuation increases, and the decrease is substantial (from .52 to .28). Furthermore, there is fair agreement between the observed and predicted results. However, the same trend that was noted in the previous example is also noticeable here, although it is much less pronounced. In the last two intervals, the empirical estimate tends to fall below the predicted estimate. Again, we conjecture that the discrepancy is because the extreme intervals also contain individuals who responded inconsistently for reasons other than large fluctuation.

The differential validity approach discussed above used the prediction error as a measure of predictability (Ghiselli, 1963). To relate the present results to this approach, we fitted a linear regression in which the E scores were taken as the predictor, and the impulsivity scores as a criterion. Next, for each individual, the prediction error was obtained as the absolute difference between his/her observed and predicted impulsivity scores. Finally, in each of the five intervals defined above, we computed the mean absolute error in each interval. The results are graphically presented in figure 5.

Table 2
Validity estimates for the five intervals on σ^2_i. Example 2

Empirical validity	.52	.38	.39	.37	.28
Predicted validity	.56	.38	.41	.41	.31

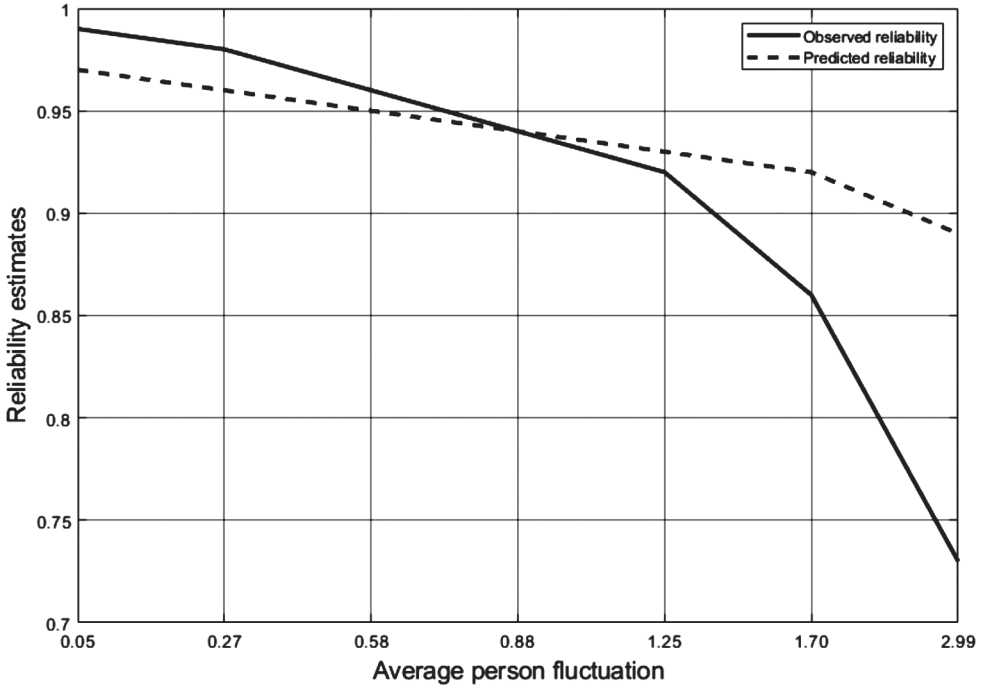

Figure 3. Reliability estimates as a function of average person fluctuation. Example 1

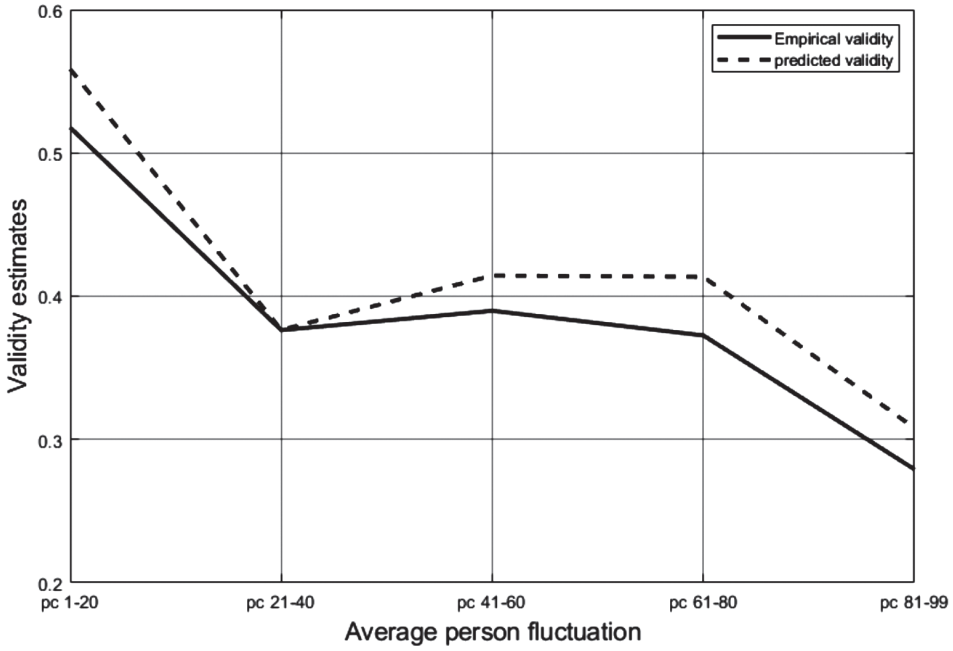

Figure 4. Validity estimates as a function of average person fluctuation. Example 2

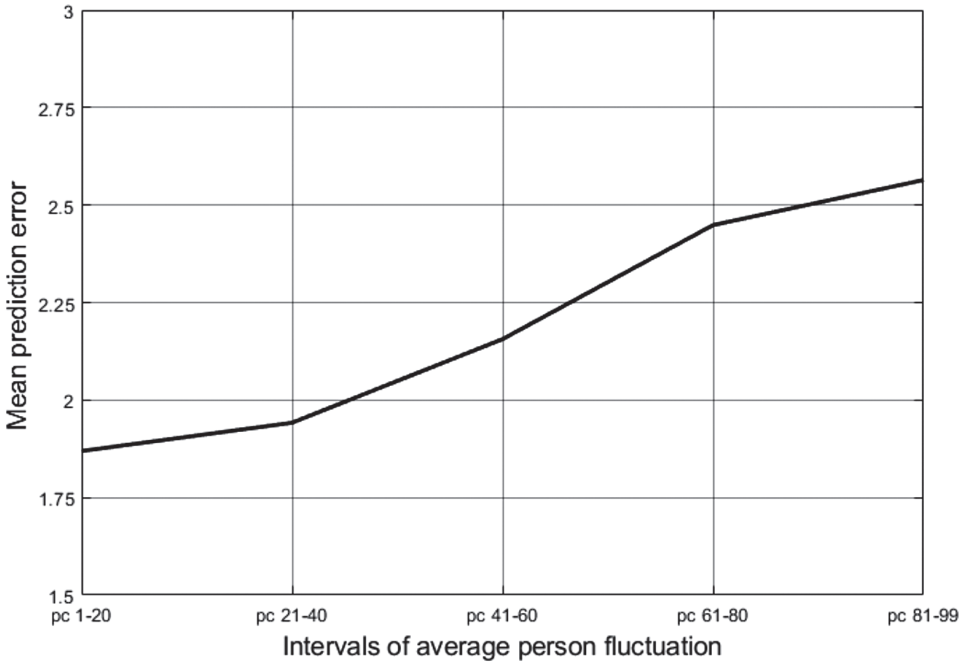

Figure 5. Mean prediction error in each interval. Example 2

The results in figure 5 are clear: prediction error increases as average fluctuation increases, as the model predicts. So, individual fluctuation can be considered to behave like a predictability index, as discussed above.

Discussion

Flexible "Dual" models in which (a) both items and persons are sources of error and (b) the amount of error generally varies over persons and over items have been considered the most appropriate for personality measurement. Recently, a formal general model of this type, as well as relatively simple and tractable procedures for fitting it have been developed. This means that issues that were considered in the personality literature, but never formally stated, can now be assessed. In particular, this paper has been concerned with the expected behavior of personality tests scores in terms of reliability and validity when both sources of error are operating.

Our study intended to serve several purposes, and we shall summarize two basic theoretical and empirical findings. First, the scores implied by the DTCRM are expected to behave in accordance with basic CTT principles in terms of reliability and validity. Second, the amount of person fluctuation is expected to place a ceiling on the amounts of empirical reliability and validity that the test scores can attain, even when the test is made from items that have small amounts of error.

We also considered that the average amounts of person fluctuation might vary across different groups of individuals, while the item characteristics remain invariant. If these assumptions hold, then both the reliability and validity of the scores in each group can be predicted as a sole function of the average group fluctuation. We checked these predictions in the two empirical examples and found substantial empirical support for them.

We consider that the findings above are relevant, and that the use of the model might be appropriate for personality measurement, provide new sources of information, and give rise to future research. At the same time, however, we note that this is only an initial study and, as such, has its share of limitations and points that deserve further research. To start with, a key issue is to appraise the relevance and the substantive implications of the ceiling effect discussed here. Assessing this issue would require intensive research based on (a) different groups of individuals expected to differ on average fluctuation (defined for example by maturity, cultural, or intellectual levels; see Navarro-González et al., 2018) and (b) different personality traits. Furthermore, these studies should be carefully controlled in psychometric terms (see Ferrando & Demestre, 2008). As an informal insight, however, our experience suggests that estimates of item reliability in (4) obtained in normal-range personality traits would be perhaps around 0.30 (they were 0.33 and 0.25 in our examples). This result translates into the 'noisiest' curve at the bottom of figure (2), and, with regards to the comments made in the introduction suggests both that there is clearly room for improving personality items, and that the impact of individual fluctuation is non-negligible. For the moment, however, evidence provided by only two empirical examples is neither strong nor generalizable enough to assume that the results predicted here will hold for personality measures in general. Furthermore, a result common to both studies suggests that some extremely inconsistent individuals who would be considered by the model to be highly fluctuating, are, in fact, responding inconsistently for reasons other than mere fluctuation (e.g. malingering, unmotivated responding, or idiosyncratic responding). So a procedure for distinguishing between different sources of inconsistency seems to be needed if the model is to lead to informative and correct predictions.

Acknowledgments

This project has been possible with the support of a grant from the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

Berdie, R. F. (1961). Intra-individual variability and predictability. *Educational and Psychological Measurement*, *21*, 663-676. http://doi.org/10.1177/001316446102100313

Coombs, C. H. (1948). A rationale for the measurement of traits in individuals. *Psychometrika*, *13*, 59-68. http://doi.org/10.1007/BF02289075

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. http://doi.org/10.1007/BF02310555

Ferrando, P. J. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology*, *9*, 150-161. http://doi.org/10.1027/1614-2241/a000060

Ferrando, P. J. (2019). A Comprehensive IRT Approach for Modeling Binary, Graded, and Continuous Responses With Error in Persons and Items. *Applied Psychological Measurement, 43*, 339-359. http://doi.org/10.1177/0146621618817779

Ferrando, P. J., & Navarro-González, D. (2020). InDisc: An R Package for Assessing Person and Item Discrimination in Typical-Response Measures. *Applied Psychological Measurement, 44*. Advance online publication. https://doi.org/10.1177/0146621620909901

Ferrando, P.J., & Demestre, J. (2008). Características de forma y contenido que predicen la capacidad discriminativa en ítems de personalidad: un análisis basado en la Teoría de Respuesta a los Ítems [Content and form characteristics that predict discriminating power in personality items: An Item Response Theory-based analysis]. *Psicothema*, *20*, 851-856.

Fiske, D. W. (1963). Homogeneity and variation in measuring personality. *American Psychologist*, *18*, 643-652. http://doi.org/10.1037/h0042883

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology, 47*, 81-86. http://doi.org/10.1037/h0047177

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment*, *89*, 167-187. http://doi.org/10.1080/00223890701468568

Hofstee W. K. B.,Ten Berge J. M. F., & Hendriks A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences*, *25*, 897-909. http://doi.org/10.1016/S0191-8869(98)00086-5

Hontangas, P. M., Leenen, I., De La Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, *28*, 76-82. https://doi.org/10.7334/psicothema2015.204

Jackson, D. N. (1986). The process of responding in personality assessment. In Angleitner, A. & Wiggins, J. S. (eds.) *Personality assessment via questionnaires* (pp. 123-142). Springer-Verlag.

LaHuis, D. M., Barnes, T., Hakoyama, S., Blackmore, C., & Hartman, M. J. (2017). Measuring traitedness with person reliability parameters. *Personality and Individual Differences*, *109*, 111-116. http://doi.org/10.1016/j.paid.2016.12.034

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694. http://doi.org/10.2466/pr0.1957.3.3.635

Lord, F. M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, *31*, 19-26. http://doi.org/10.1111/j.2044-8317.1978.tb00568.x

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, *4*, 1-7. http://doi.org/10.1177/014662168000400101

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223-237. http://doi.org/10.1207/s15327906mbr2903_2

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683-729. http://doi.org/10.1111/j.1744-6570.2007.00089.x

Mosier, C.I. (1942). Psychophysics and mental test theory II: The constant process. *Psychological Review*, *48*, 235-249.

Muñiz, J. (2000). *Teoría Clásica de los Tests* [Classic Test Theory]. Pirámide.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*, 7-16. https://doi.org/10.7334/psicothema2018.291

Navarro-González, D., & Ferrando P. J. (2020). InDisc: Obtaining and Estimating Unidimensional IRT Dual Models. R package version 1.0.3. https://CRAN.R-project.org/package=InDisc

Navarro-González, D., Ferrando, P. J., & Vigil-Colet, A. (2018). Is general intelligence responsible for differences in individual reliability in personality measures? *Personality and Individual Differences*, *130*, 1-5. https://doi.org/10.1016/j.paid.2018.03.034

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92*, 486-511. http://doi.org/10.1037/0033-295X.92.4.486

Sireci, S., & Padilla, J. L. (2014). Validating assessments: Introduction to the Special Section. *Psicothema*, *26*, 97-100. https://doi.org/10.7334/psicothema2013.255

Taylor, J. B. (1977). Item homogeneity, scale reliability, and the self-concept hypothesis. *Educational and Psychological Measurement*, *37*, 349-361. http://doi.org/10.1177/001316447703700209

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-278.