SECCIÓN METODOLÓGICA

Psicológica (2007), 28, 237-257.

The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT

Pere J. Ferrando♦ and Eliseo Chico

'Rovira i Virgili' University (Spain)

A theoretical advantage of item response theory (IRT) models is that trait estimates based on these models provide more test information than any other type of test score. It is still unclear, however, whether using IRT trait estimates improves external validity results in comparison with the results that can be obtained by using simple raw scores. This paper discusses some methodological results based on the 2-parameter logistic model (2PLM) and is concerned with three issues: first, how validity coefficients based on IRT trait estimates must be interpreted; second, how inferences about these coefficients can be made; and third, which differences in external validity can be expected if the 2PLM is correct for the data and IRT scores are used in place of raw scores. Four empirical examples in the personality domain provided further evidence for the results that can be expected in real research in which the model is, at best, a good approximation to the data. A general result of these examples was that validity coefficients based on IRT scores were similar to those based on raw scores.

One of the theoretical advantages of item response theory (IRT) over classical test theory (CTT) is greater accuracy in the estimation of the individual trait levels. For most standard IRT models, using maximum likelihood (ML) trait estimates obtained from the patterns of item responses provides more test information (and therefore greater accuracy) than using any unweighted or weighted type of test score (Birnbaum, 1968; Samejima, 1969). Indeed, this advantage applies only if the IRT model is correct for the data.

♦ Acknowledgments: This research was partially supported by a grant from the Spanish Ministry of Science and Technology (SEJ2005-09170-C04-04/PSIC) with the collaboration of the European Fund for the Development of Regions. Correspondence should be addressed to: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicología. Carretera Valls s/n. 43007 Tarragona (SPAIN). E-mail: perejoan.ferrando@urv.cat

If IRT-based ML trait estimates are more accurate than the common test scores used in CTT, it is reasonable to expect that using these estimates would lead to improved relations with relevant external variables (Lumsden, 1976). Broadly speaking, as external variables we consider here either non-test variables or scores in other tests. In modern validity theory, these types of relations are considered external components of validity (Messick, 1993). Here we will use the shorter expression 'external validity' to refer to this source of validity evidence.

At present, the conjecture described above does not appear to have been studied systematically. Hambleton and Swaminathan (1985) and Lumsden (1976) complained about the lack of validity studies based on IRT scores and considered that they were very necessary. Despite this advice, however, a search of the literature suggests that external validity based on IRT estimates is a rather neglected topic. There appear to be no methodologically oriented studies that discuss the results that the theory suggests, and empirical studies (based on real or simulated data) comparing IRT and CTT trait estimates in terms of validity are very scarce. Several previous studies compared IRT and CTT but their comparisons mainly focused on item and person statistics (e.g. Fan, 1998; Lawson, 1991; MacDonald & Pauonen, 2002). The results of these studies suggest that CTT and IRT person estimates correlate very highly. The few documented empirical studies on external validity either did not find practical improvements in the validity coefficients (Ferrando, 1999; Young, 1995), or found slight improvements only under certain conditions. Thus, McBride and Martin (1983) found that the validity coefficients of IRT scores were slightly better than those of raw scores for short tests up to 10 items, but above this there were no practical differences. In a second study, they found improvements in the IRT-based validity coefficients that tended to level-off at around 30 items (McBride & Martin, 1983).

The aim of this paper is to discuss some results concerning the external validity of ML scores based on the two-parameter logistic model (2PLM) and to compare the expected validity coefficients based on either the ML estimates or the simple raw scores. To do so, we derive some validity results from the principles and assumptions of the model. These derived results are not new, and most of them (or fairly similar ones) have previously appeared in the literature, though generally related to topics other than validity. The main aim of this paper is not, therefore, to present innovative methodology but, aimed at the applied researcher, to highlight and discuss three main points: (a) how empirical validity coefficients based on ML trait estimates should be interpreted; (b) how inferences about the disattenuated coefficients based on the ML estimates can be made, and (c) what validity differences (if any) should be expected in real applications if

ML estimates are used in place of simple raw scores. The results and procedures we discuss are complemented by four empirical examples in personality measurement. These examples illustrate points (a) and (b) and provide additional information about point (c). The 2PLM was chosen because it is the model we use most in our applied research (personality measurement) to analyse questionnaires made up of binary items. However, our results also apply to the simpler one-parameter model. Indeed, similar results could be derived for the three-parameter model but we will not discuss that model here.

As far as point (c) above is concerned, some authors suggested that no practical improvements in external validity should be expected if IRT scores based on the models above are used in place of raw scores (Nunnally, 1987, Reise, 1999, Schmidt & Hunter, 1999). There are both theoretical and practical bases for this prediction. At the theoretical level, the raw score is a sufficient statistic for the one-parameter model whereas for the 2PLM the sufficient statistic is a weighted sum of the item scores. In more detail, for the one-parameter model the optimal scoring weights are equal weights, whereas for the 2PLM they are the item discriminations (e.g. Lord & Novick,1968). Optimally weighted linear composites are just as good measures of trait level as IRT trait estimates, so not a great many differences in validity are expected if some scores are used instead of others. This rationale, however, needs to be qualified. IRT scores and linear-composite scores are nonlinearly related and, given that the usual validity coefficients are correlation coefficients, the use of one type of score or another can give rise to clear differences in validity, particularly if there are strong end (floor or ceiling) effects.

At the practical level, certain authors argue that no substantial validity improvements can be expected, because, although IRT and raw scores are nonlinearly related, they generally correlate very high (above .9). This, however, should also be qualified. First, even with a correlation above .9 there is still room for improvement in the validity coefficient (Stanley & Wang, 1969). Second, we believe that the conditions in which IRT and raw scores are expected to be more or less related, and how this can affect the external validity coefficients, should be studied in greater depth.

The External Validity of ML Trait Estimates

Consider an external non-test variable y that is linearly related to a trait θ, the trait scaled with zero mean and unit variance. Using Lord and Novick's (1968) distinction, we shall define the theoretical validity coefficient as $\rho_{y\theta}$, the product-moment correlation between θ and the external variable.

Now consider a test made up of n binary items that measures θ, so that the item responses behave according to the two-parameter logistic model (2PLM):

$$P_j(\theta) = P(X_j = 1 \mid \theta) = \frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}}, \quad (1)$$

where D=1.702 is a scaling constant, a_j is the discrimination parameter of item j, and b_j is the difficulty or location parameter or item j. The item parameters are assumed to be fixed and known values and are used in the scoring phase to obtain maximum likelihood (ML) estimates based on the patterns of item responses (see e.g. Mislevy & Bock, 1990). For a respondent i, the ML estimate can be written as:

$$\hat{\theta}_i = \theta_i + \omega_i, \quad (2)$$

where θ_i is the 'true' trait level and ω_i is the error of measurement. Asymptotically (as the number of items increases without limit), the ML estimate is conditionally unbiased and the conditional variance is (see e.g. Hambleton & Swaminathan, 1985):

$$Var(\hat{\theta}_i \mid \theta_i) = \frac{1}{I(\theta_i)}, \quad (3)$$

where $I(\theta)$ is the test information function, which for the 2PLM is given by

$$I(\theta) = \sum_{j=1}^{n} D^2 a_j^2 P_j(\theta)(1 - P_j(\theta)). \quad (4)$$

The marginal mean and variance of the ML estimate can be obtained by using the corresponding decomposition formulas

$$E(\hat{\theta}) = E_\theta(E(\hat{\theta} \mid \theta)) = E(\theta) = 0, \quad (5)$$

and

$$Var(\hat{\theta}) = E_\theta(Var(\hat{\theta} \mid \theta)) + Var_\theta(E(\hat{\theta} \mid \theta)) =$$
$$= E_\theta(\frac{1}{I(\theta)}) + Var(\theta) = E_\theta(\frac{1}{I(\theta)}) + 1 \qquad (6)$$

We shall now define the empirical validity coefficient as the product-moment correlation between y and the ML estimates of θ. From the assumptions so far stated, and using results (5) and (6) together with standard covariance algebra, the empirical validity coefficient is found to be

$$\rho_{y\hat{\theta}} = \rho_{y\theta} \sqrt{\frac{1}{1 + E_\theta(\frac{1}{I(\theta)})}}. \qquad (7)$$

Equation 7 shows that the empirical coefficient $\rho_{y\hat{\theta}}$ is an attenuated (downwardly biased) estimate of the theoretical coefficient $\rho_{y\theta}$, and that the squared root term on the right hand side of (7) is the attenuation factor. To see more explicitly that (7) is an attenuation formula, we use the standard definition of the reliability of ML scores as a variance ratio (e.g. Mislevy & Bock, 1990; Mellenbergh, 1996; Samejima, 1994).

$$\rho_{\hat{\theta}\,\hat{\theta}} = \frac{Var(\theta)}{Var(\hat{\theta})} = \frac{1}{1 + E_\theta(\frac{1}{I(\theta)})}. \qquad (8)$$

Equation 8 can be considered as an approximation reliability formula obtained from the asymptotic properties of the ML estimator. Bechger, Maris, Verstralen & Béguin (2003) provided a more general framework for the reliability of any trait estimate, and showed that equation 8 can be considered as an upper limit for the reliability.

Using now equation 8 in equation 7 we obtain

$$\rho_{y\theta} = \frac{\rho_{y\hat{\theta}}}{\sqrt{\rho_{\hat{\theta}\hat{\theta}}}}, \tag{9}$$

which has the same form as the correction-for-attenuation formula in CTT when the external variable (the criterion) is unaltered.

We shall now discuss the factors that determine the magnitude of the attenuation of $\rho_{y\hat{\theta}}$ with respect to $\rho_{y\theta}$. First, from equation 4, the average test information is found to be

$$E_\theta(I(\theta\ )) = \sum_{j=1}^{n} D^2 a_j^2 E_\theta(P_j(\theta)(1 - P_j(\theta))), \tag{10}$$

where $E_\theta(I(\theta))$ denotes the expectation over the different θ levels. Two factors are explicit in equation 10: the number of items and the magnitudes of the item discriminatory powers. Therefore, the average information increases (and the amount of attenuation decreases) as the test becomes longer and the items become more discriminating. Note that these factors are the same factors that determine the amount of attenuation in the CTT-based correction-for-attenuation formula. The role of the item locations, on the other hand, is not so immediate and, to assess this point, some specific distribution must be assumed for θ. In IRT applications, θ is generally assumed to be normally distributed. If this is so, and the other two factors remain constant, the maximum average information (and, therefore, the minimum attenuation) is attained when all the items have locations of zero (the mean of θ). The explanation for this is as follows. For a given trait level, an item score provides maximal information when the item location matches the trait level. If the trait is normally distributed, most of the respondents are concentrated around the mean, so the average test information will be maximal when all of the items are located near the mean.

We shall now summarise and discuss the results so far described from an applied point of view. IRT-based ML scores (or trait estimates) are fallible measures of the 'true' trait levels. Therefore, the empirical validity coefficient is an attenuated estimate of the theoretical validity coefficient. The amount of attenuation decreases as the test becomes longer and the

items become more discriminating. Moreover, if the distribution of the trait is bell-shaped and symmetrical, the amount of attenuation will be reduced if the items are located near the trait mean.

Commercially available IRT programs such as BILOG provide the average value of the test information or the average reliability as part of the output. These values can be used in equations 8 or 9 to obtain a disattenuated estimate of the theoretical validity coefficient. As in CTT, the estimated disattenuated validity coefficients might be meaningful and add interesting information in certain research scenarios. However, we have been unable to find any applied IRT-based validity study in which disattenuated coefficients have been obtained.

The correction-for-attenuation formula can be directly extended to the case in which validity analysis involves relations between two sets of scores. Consider two tests that measure related traits, θ_1 and θ_2. The tests are made up of binary items, and both can be fitted by the 2PLM. The relation between the theoretical validity coefficient $\rho_{\theta_1\theta_2}$ and the empirical validity coefficient $\rho_{\hat{\theta}_1\hat{\theta}_2}$ is given by

$$\rho_{\theta_1\theta_2} = \frac{\rho_{\hat{\theta}_1\hat{\theta}_2}}{\sqrt{\rho_{\hat{\theta}_1\hat{\theta}_1}}\sqrt{\rho_{\hat{\theta}_2\hat{\theta}_2}}}. \tag{11}$$

Bechger et al. (2003, Theorem 3) derived equation 11 (and so 9) by using a more general framework, and showed that the result in equation 11 applies to any pair of unbiased estimates with uncorrelated measurement errors. The interested reader is referred to this paper for a detailed derivation of these equations.

Overall, the results so far discussed were obtained by assuming that the 2PLM was the correct model for the data and by using asymptotic properties of the ML estimator. With regard to the first point, strictly speaking, the present results apply only if the model is correct. However, models never fit the data perfectly and are at best only simplified approximations to reality. In practical applications, both the model assumptions and the goodness of model-data fit must be carefully assessed (see Hambleton & Swaminathan, 1985, Chapter 8) but, even if the results are positive, we can only say that the model provides a reasonably good approximation to the data. With regard to the second point, for tests of finite length the ML estimator is not unbiased, and the reciprocal of the test information is not exactly the conditional variance in (3). However, it has

been found that, in practice, the asymptotic properties of ML are reasonably well met, even for tests of moderate size, provided that the amount of information is large enough (Samejima, 1977, 1994). Corrections for both the bias of the ML estimator and the information function are available (Lord, 1983; Samejima, 1994), and could be used in equations 7 and 9 to obtain theoretically more accurate estimates of the disattenuated validity. These corrections, however, are very complex and it is not clear that are necessary in practical applications.

Inferences about the Disattenuated Coefficients

The results discussed in the preceding section are population-based, and the sole source of error considered for $\rho_{y\hat{\theta}}$ is measurement error (unreliability). Most real studies, however, are sample based. So the empirical validity coefficient $r_{y\hat{\theta}}$ is affected by: (a) unreliability and (b) sampling variability. In the previous section, we described an IRT-based correction-for-attenuation formula which has the same form as the corresponding CTT-based formula, and which deals with the first source of error. In this section we shall discuss procedures for addressing the second source of error: sampling variability. More specifically we shall show how a covariance structure model, which was developed for making inferences about the CTT-based disattenuated coefficient, can be readily adapted for making inferences about the IRT-based disattenuated coefficient. In particular, we shall consider three types of inference: (a) obtaining confidence intervals, (b) testing the hypothesis that the theoretical validity coefficient is zero in the population of interest, and (c) testing the hypothesis that the theoretical validity coefficient is unity. Situation (c) is mainly concerned with validity studies in which two sets of test scores are supposed to measure the same trait.

The original model for making inferences about the CTT-based disattenuated correlations was developed by Hancock (1997). The model adapted for making inferences about the IRT-based disattenuated validity coefficient in equations 9 and 11 is depicted in Figure 1 for the case of an external non-test variable (i.e. equation 9).

In the model in Figure 1, the variances of θ and θy are fixed to unity, the indicator paths are fixed at the values given in the figure and the error paths are left free. Note that only the ML estimates are corrected for unreliability. With the above constraints, the free covariance to be estimated, $\sigma_{y\theta}$, is precisely the disattenuated correlation value that would be obtained by using equation 7. The free parameters of the model are

estimated from the 2×2 covariance matrix between the ML estimates and the external variable, and this makes the model just identified. However, fitting the model provides a standard error for the $\sigma_{y\theta}$ parameter estimate.

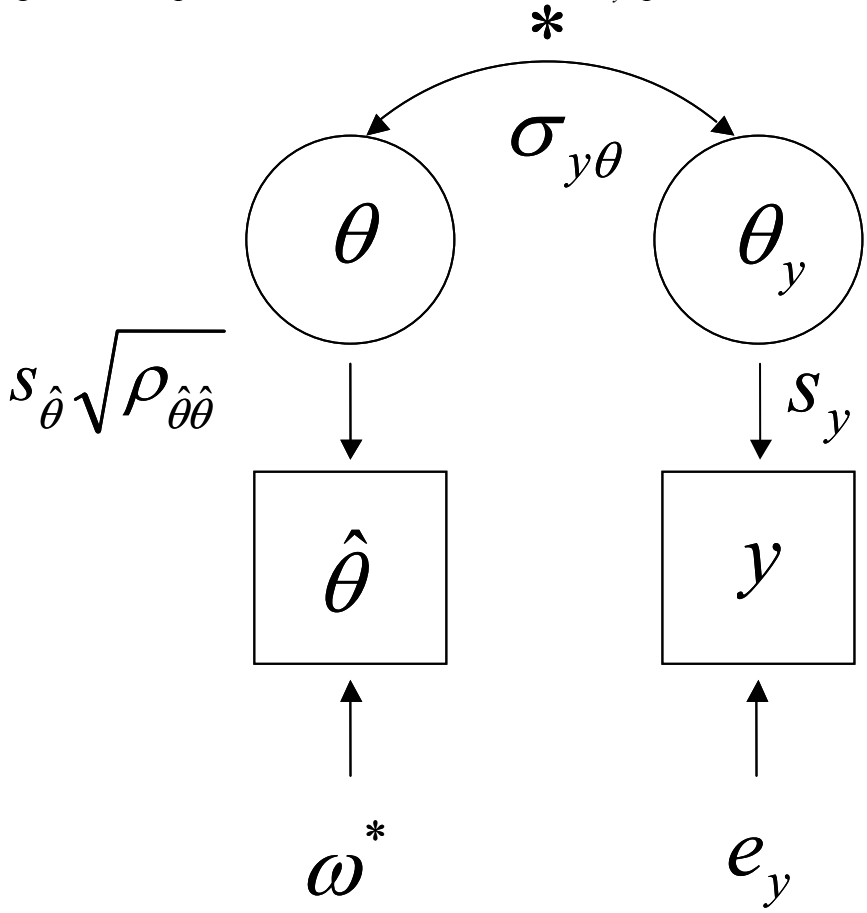

Figure1. Model for Making Inferences about the Disattenuated Validity Coefficient.

For non-extreme estimated values, this standard error can be used to construct an approximate confidence interval based on the normal distribution. If we further impose the restriction that $\sigma_{y\theta}$ is zero or unity, the model becomes overidentified and hypotheses (b) and (c) above can be tested using a goodness-of-fit statistic with one degree of freedom. Finally, the model can be extended to assess relations between two sets of scores. In this case, both paths would be corrected for unreliability and the covariance to be estimated in the least restricted model would be precisely the disattenuated correlation given in equation 11.

The model and procedures so far described are expected to be useful for making inferences about validity results based on IRT trait estimates. However, strictly speaking, they are only approximate. Fixing the path indicator to the value in figure 1 implies treating the standard deviation of the ML estimates and the reliability estimates as if they were fixed and known values, while, in fact, both are also subject to sampling variability. By standard sampling theory, the approximation is expected to improve the larger and more representative the sample is.

The External Validity of Raw Scores

In this section we assume that the 2PLM is correct for a given set of data, and discuss some results concerning the external validity of the usual raw test scores, which are obtained as the simple sum of item scores: $X=X_1+X_2+\ldots+X_n$.

The relations are now more complex than in the first section. According to equation 2, the ML estimate is linearly related to θ. However, according to the 2PLM, the relation between X and θ is nonlinear. A model-based expression for the empirical validity coefficient ρ_{YX} can be obtained by considering the 2PLM as an approximation to the normal ogive model and by extending some of Lord's (1952, 1953) results in order to incorporate the external variable. The resulting expression, however, is complex and does not lead to clear interpretations (Ferrando, 2004).

A clearer, more interpretable result can be obtained by considering a simple situation based on 'parallel' items, all of which have the same discrimination ($a_j=a$) and difficulties of $b_j=0$ (note that this is the condition that minimises attenuation in the case of ML estimates). Under these conditions, the model-based empirical validity coefficient is found to be

$$\rho_{yx} = \rho_{y\theta}\sqrt{\frac{1}{1+\frac{1}{n}\left(\frac{(1+a^2)}{0.25D^2a^2}-1\right)}}. \qquad (12)$$

It can be shown that equation 12 is also approximately correct for items with difficulties distributed uniformly with a mean of zero, and with different discrimination values (in this case the average of the a_j's would be used in place of the constant a value).

Equation 12 has the same form as (7) and is a correction for attenuation formula. For fixed item locations, the other factors that

determine the amount of attenuation are the same as in (7): the number of items n, and the magnitude of the item discriminations. It therefore follows that, if the 2PLM is correct, both empirical validity coefficients (7) and (12) are asymptotically unbiased and converge to the theoretical validity coefficient as the number of items increases without limit.

We have been unable to find a simple expression to relate the attenuation factors (the terms within the square root) in equations 7 and 12. However, the respective amounts of attenuation can be obtained and tabled for different values of test length and average discrimination. This provides an idea of which validity results can be expected in applied research when one type of test scores or another is used. These results are shown in Table 1. The attenuation in (7) was obtained by assuming that the distribution of θ was standard normal.

Table 1. Predicted Values of the Attenuation Factor for ML Estimates (Upper values) and for Raw Scores (Lower Values) Under Different Conditions of Test Length and Item Discrimination.

	a=0.25	a=0.50	a=0.75	a=1.00	a=1.25
n=10	0.55 0.53	0.78 0.77	0.87 0.86	0.91 0.91	0.93 0.93
n=20	0.68 0.66	0.87 0.86	0.93 0.93	0.95 0.95	0.96 0.96
n=30	0.75 0.74	0.91 0.90	0.95 0.95	0.97 0.97	0.97 0.97
n=40	0.79 0.78	0.93 0.92	0.96 0.96	0.97 0.97	0.98 0.98

As an example of how the table is used, consider a test of 20 items with difficulties centred around the trait mean (or uniformly distributed around this value) and with an average discrimination of 0.50. Suppose that the theoretical validity coefficient is 0.40. The empirical validity coefficient based on the ML estimates is then expected to be: $0.40 \times 0.87 = 0.348$. The empirical validity coefficient based on the raw scores is expected to be: $0.40 \times 0.86 = 0.344$.

Table 1 shows that, as expected, the attenuation decreases in both cases as the test becomes longer and the items become more discriminating. The main result, however, is that the expected differences between both

types of scores are minimal, and are only appreciable in the case of short tests with poorly discriminating items. For well-designed tests of the type that we expect to find in applied studies, Table 1 suggests that the expected validity differences due to the use of one type of scores or another are negligible.

We should again note that the results in Table 1 can only be considered to be approximations and are based on a series of assumptions concerning the item parameters and the distribution of the trait. Most important, however, is that the results are obtained by assuming that the 2PLM is the correct model for the data of interest. As we discussed above, models are never correct and we only can assess whether the assumptions of the model are reasonably fulfilled and whether the model is a reasonably good approximation to the data. The empirical studies below aim to provide additional guidance about the results that can be expected in real applications when the 2PLM provides an appropriate (but not perfect) fit to the data.

Empirical Examples

Below we discuss four empirical studies based on data we collected in our research domain, which is personality measurement. The first two studies are concerned with relations to external non-test variables. The last two studies are concerned with relations to scores on other tests. In all cases, the tests were based on binary items and fitted by the 2PLM using a two-step procedure based on marginal ML estimation as implemented in the BILOG program (Zimowski, Muraki, Mislevy & Bock, 2002). First, the items were calibrated and the assumptions of the model (unidimensionality) as well as the model-data goodness of fit were assessed. Second, the item parameters were taken as fixed and known values and used to obtain ML estimates for each response pattern. In all the examples discussed, the unidimensionality condition was reasonably fulfilled, and the 2PLM provided appropriate fit to the data. Below we discuss only results related to the validity issues considered in this paper. Details of the assessment of model-data fit can be obtained from the authors.

In all cases, inferences about the disattenuated correlations based on Hancock's modified structural model were based on ML estimation as implemented in the LISREL program (Jöreskog & Sörbom, 1996).

Study 1: Anxiety and Academic Performance

The Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1978) is a multidimensional questionnaire that measures different aspects of anxiety in schoolchildren. In previous studies we found that the sub-scales of the RCMAS were well fitted by the 2PLM and that the 7-item Concentration subscale scores were the ones that were most clearly related to different measures of academic performance. In the present example, we used data obtained in a sample of 1022 Spanish schoolchildren between 11 and 12 years of age. As external variables, we considered three measures of academic performance. These were the final marks in Language, Mathematics and Social Sciences.

The results of the item calibration of the Concentration subscale showed that the average item discrimination was 0.78, which is fairly high for personality items. The average information (see equation 10) was 2.10, and the corresponding marginal reliability (see equation 8) was 0.68, which can be considered acceptable for such a short test. The item difficulties were evenly distributed around the trait mean. The validity results based on both the ML scores and the raw scores are shown in Table 2.

Table 2. Summary of Validity Results for Example-1 data: Anxiety and Academic Performance.

External variable	$r_{\hat{\theta}y}$	r_{XY}	$\hat{\rho}_{\theta y}$	*90% C.I.*	$\chi^2(1)$
Language	-.25	-.26	-.31	(-.37;-.24)	69.04
Mathematics	-.30	-.32	-.37	(-.43;-.31)	99.80
Social Sciences	-.30	-.31	-.38	(-.44;-.32)	104.95

Note. $r_{\hat{\theta}y}$ =empirical validity coefficient based on the ML estimates; r_{XY} =empirical validity coefficient based on the raw scores ; $\hat{\rho}_{\theta y}$ and 90% C.I =point and 90% confidence interval estimates of the disattenuated correlation based on the ML scores; $\chi^2(1)$= goodness-of-fit statistic for testing the null hypothesis $\rho_{\theta y}=0$.

The empirical validity coefficients are quite acceptable if we consider that the predictor of academic performance is not a measure of ability but a measure of anxiety. The estimated theoretical coefficients (i.e., disattenuated correlations) are clearly higher than the usual values found in personality measurement. In this example, Hancock's modified model was used to obtain confidence interval estimates of the disattenuated validity coefficients and to test the hypothesis that these coefficients were zero in

the population. In all cases the hypothesis is clearly rejected. This result also follows if we inspect the confidence intervals: the zero value falls outside the confidence interval and quite far from the upper limit. Finally, the empirical validity coefficients based on the ML scores are quite similar to those based on the raw scores (though the coefficients based on the raw scores are slightly higher in absolute value). The observed values of these validity coefficients agree quite closely with the predictions that can be made from Table 1 based on the number of items and the average item discrimination.

Study 2: Psychopathy and Criminal Behaviour

In the second example, Hare's Psychopathy Checklist Revised (PLC-R; Hare, 1991) was administered to a group of 298 male prisoners convicted of various offences. The PLC-R is essentially a measure of antisocial behaviour, but previous research suggests that its scores are also related to different external measures of criminal behaviour (Cooke, 1995; Hare, 1970). The PLC-R version used in this study consisted of 20 binary items and was completed by the prison psychologist in a semistructured interview. As external measures we used (a) the number of crimes, and (b) a measure of the type and gravity of the crime: 0 non-violent, 1 partly violent, and 3 violent.

In the item calibration phase, the average item discrimination was estimated to be 0.75, which is fairly high for a personality measure. The average item difficulty was 1.04. The item locations were therefore not centred around the trait mean but located rather more to the right. In other words, the test was 'difficult' even for this population. The average information was 5.54 and the corresponding marginal reliability was 0.85, which is fairly acceptable. The validity results are shown in Table 3.

The empirical validity coefficients were much higher than is usual in personality measurement, and the disattenuated coefficients and confidence intervals clearly showed that the hypothesis that these coefficients are zero in the population was untenable in both cases. With regard to comparisons and predictions, the empirical validity coefficients based on the ML scores and on the raw scores were again quite similar. Moreover, the observed values of these validity coefficients agreed closely with the predictions that can be made from Table 1 based on the number of items and the average item discrimination. Note that the table predictions are accurate even in this case, where the item difficulties do not meet the assumptions on which the table was based (the difficulties were not centred around the trait mean).

Table 3. Summary of Validity Results for Example-2 Data: Psychopathy and Criminal Behaviour.

External variable	$r_{\hat{\theta}y}$	r_{XY}	$\hat{\rho}_{\theta y}$	*90% C.I.*
nº offences	.40	.38	.44	(.37;.51)
Type of offence	.50	.51	.54	(.47;.61)

Note. $r_{\hat{\theta}y}$ =empirical validity coefficient based on the ML estimates; r_{XY} =empirical validity coefficient based on the raw scores $\hat{\rho}_{\theta y}$ and 90% C.I =point and 90% confidence interval estimates of the disattenuated correlation based on the ML scores.

Study 3: Impulsivity and Sociability

In Eysenck's personality system, Extraversion is conceived as a broad, higher-order factor comprising several correlated primary factors. The most well identified primary factors are 'Impulsivity' and 'Sociability'. The EPI-A Estraversion scale (Eysenck & Eysenck, 1963) consists of items that can be allocated into an 'Impulsivity' subscale and a 'Sociability' subscale (Eysenck & Eysenck, 1963; Rocklin & Revelle, 1981). Whether these subscales need to be scored separately or form part of the general Extraversion scale is controversial (Carrigan, 1960). From a methodological point of view, it may be interesting to study the disattenuated correlation between the scores obtained in both subscales. This would provide additional information about the appropriateness of scoring them separately or not.

In the present example, the EPI-A was administered to a sample of 335 undergraduate students and the 10 Impulsivity items and 13 Sociability items were calibrated separately using the 2PLM. The results of the item calibration phase are shown in Table 4. In both cases the distribution of the item difficulties is fairly centred around the trait mean. On average, the items of the impulsivity subscale are more discriminating than the items of the sociability subscale. The reliabilities based on the average information are acceptable in both cases.

The empirical validity coefficient based on the ML estimates was 0.39 and the corresponding coefficient based on the raw scores was 0.40. As expected from the theory, these estimates were practically the same. The disattenuated coefficient based on the ML estimates was 0.50 and the (approximate) 90% confidence interval was (0.37; 0.63). The restriction that the disattenuated coefficient was unity gave a chi-squared value of 26.99 with one degree of freedom. It appears, therefore, that the two sets of scores

are clearly related but that the relation is far from perfect. Indeed, the decision to consider the scores separately or as part of a unitary construct must also be guided by theoretical considerations, but these results suggest it is reasonable to treat the scales separately.

Table 4. Summary of Item and Trait Estimates for Example-3 data: Impulsivity and Sociability.

	Impulsivity sub-scale	Sociability sub-scale
Average *a*	.94	.70
Average *b*	.37	-.24
$E(I(\theta))$	3.79	3.28
$\rho_{\hat{\theta}\hat{\theta}}$	.79	.77

Study 4: Lies and Social Desirability

Personality questionnaires tend to include a Lie scale to detect faking. However, when the tests are administered in neutral (low motivating) conditions, the Lie scores are assumed to measure a consistent trait which has been labelled as a 'need for approval' or 'social conformity' (Furnham, 1986). This trait is thought to be the same as that measured by standard social desirability (SD) scales, which are made up of items whose responses can place the individual in a positive light.

In the present example, the Lie scale of the EPQ-R (Eysenck, Eysenck & Barrett, 1985) and Crowne and Marlowe's (1960) SD scale were administered voluntarily and anonymously to a sample of 489 undergraduate students. As in the previous example, the hypothesis of most interest here is whether the theoretical (disattenuated) validity coefficient is unity, i.e., whether both scale scores measure the same trait.

The results of the item calibration phase are given in Table 5. In both cases the distribution of the item difficulties is fairly centred around the trait mean. The scales are longer than in the previous examples (21 and 33 items), but the item discriminations, especially in the SD scale, tend to be lower. The reliabilities based on the average information, however, are still acceptable.

We now turn to the validity results. The empirical validity coefficient based on the ML estimates was 0.70, and the corresponding coefficient based on the raw scores was 0.67. Again, as expected, they are quite

similar, though in this case the ML-based coefficient is a little higher. The disattenuated coefficient based on the ML estimates was 0.83, and the (approximate) 90% confidence interval was (0.72; 0.94).

Table 5. Summary of Item and Trait estimates for Example-4 data: Lie and Social Desirability.

	Lie scale	Crowne-Marlowe SD scale
Average *a*	.70	.58
Average *b*	.10	.17
$E(I(\theta))$	4.84	5.57
$\rho_{\hat{\theta}\theta}$	.83	.85

The restriction that the disattenuated coefficient was unity gave a chi-squared value of 5.60 with one degree of freedom. Strict adherence to the statistical test would lead to rejection of the null hypothesis (p=0.02). With a sample of almost 500 participants, however, the test is very powerful. Additional measures of model fit were: RMSEA=0.09, GFI=0.99 and NNFI=0.98. Overall, we believe it is reasonable to consider that both sets of scores essentially measure the same trait.

Discussion

The present paper makes mainly two types of contributions. First we interpret validity coefficients based on IRT trait estimates and show how inferences can be made about these coefficients. Second we compare the external validity results theoretically and empirically based either on trait estimates or simple raw scores. Our results do not refer to IRT models in general but to one of the most widely used models in applied research: the two-parameter logistic model. These results also apply to the popular one-parameter logistic model.

With regard to the first contribution, this paper shows that IRT trait estimates are fallible and affected by both measurement error and sampling variability. Taking measurement error first, as in CTT the empirical validity coefficient based on the trait estimates is shown to be an attenuated estimate of the theoretical validity coefficient. In this paper we have discussed the conditions that lead to a greater or lesser amount of attenuation, and obtained a correction-for-attenuation formula that has the same general

form as the well-known Spearman-CTT formula. The conditions that minimise the attenuation effect agree with conventional wisdom in psychometrics: test length, item discrimination values and distribution of item difficulties. This result has implications for test construction. If maximising external validity is the main concern, the test should be a long one and have highly discriminating items, and difficulties should either be located around the trait mean or distributed uniformly around it. This last condition is based on the assumption that the distribution of the trait is bell-shaped and symmetrical. Now turning to sampling variability, we have shown that Hancock's (1997) covariance structure model can be readily adapted to make inferences about validity coefficients based on IRT trait estimates.

With regard to the second contribution, this paper shows that, under the conditions assumed, no practical differences are expected if the simple raw scores are used instead of the ML trait estimates. Indeed this result also applies to linear transformations of the raw scores. At first sight, this result may show that, once again, one of the theoretical advantages of IRT does not result in practical improvements. However, this interpretation is too simplistic and we believe the issue deserves further consideration.

These results do not imply that raw scores will generally lead to the same results as ML estimates but that, if the IRT model is correct—or approximately correct—the validity results are expected to be virtually the same with one type of scores or another. The point here is that, to a certain extent, the IRT model is falsifiable and that both its assumptions unidimensionality (as in this case) and appropriateness can and must be tested by a goodness-of-fit investigation. If the model is considered appropriate, and external validity is the main research concern, no improvements should be expected by using more sophisticated trait estimates. On the other hand, the present results imply that raw scores obtained from a CTT approach would lead to correct validity inferences provided that the response model considered here was tenable. However, since CTT is mainly descriptive, we cannot assess whether the model is correct or not, so we do not know whether the CTT validity inferences based on the raw scores are correct.

The above discussion notwithstanding, it is disappointing that using more accurate scores does not lead to improvements in validity, even in theory. Again this result requires further consideration. Mellenbergh (1996) differentiated between two aspects of measurement precision: conditional precision and unconditional precision. Conditional precision is precision for a given respondent and is related to the concept of information. Unconditional precision is the precision in a population of respondents and

is related to the concept of reliability. The validity developments considered in this paper are concerned with unconditional precision and it is here that no improvements in validity are expected. More specifically, when estimating validity coefficients, the information is averaged over the population of respondents, and the result that this information is generally different at different trait levels (i.e. conditional precision) is not taken into account. So, in spite of the results at the unconditional level, the estimates obtained from the complete pattern of responses might still be superior to raw scores for making accurate inferences at the individual level (for example, making decisions based on a cut-off point). Again, this is a theoretical advantage of IRT that should be assessed at the empirical level.

As with any study, this paper has several limitations. The theoretical results were obtained under simple, ideal conditions. For example, the ML trait estimates are asymptotically unbiased but they are biased for tests of finite length. This would require corrections that are far more complex than the simple results used here (Lord, 1983; Samejima, 1994). Also, the distribution of the trait levels was assumed to be normal and the item difficulties were assumed to be centred around the trait mean. However, there are tests in which these conditions cannot be reasonably expected (for example, tests used for selection around a cut-off point). Another theoretical limitation is that the results do not apply to IRT models in general but only to a particular model. As far as the empirical studies are concerned, they are not intended to be exhaustive but to provide examples. At best, they provide an idea of what can be expected in personality measurement. To make these findings more generalizable, therefore, further studies are needed that have greater theoretical sophistication and take into account different trait and item distributions, different models (graded response models, for example) and different measurement domains (ability and attitude, etc.). Despite the obvious limitations of this paper, however, we suspect that any results obtained in further studies would essentially be the same as those presented here.

RESUMEN

La validez externa de las puntuaciones estimadas mediante el modelo logístico de dos parámetros: Algunas comparaciones entre la TRI y la TCT. Una de las ventajas teóricas de los modelos de la teoría de respuesta al ítem (TRI) es que las estimaciones de los niveles en el rasgo son más precisas (en términos de información) que cualquier otro tipo de puntuación. Sin embargo, no está claro aún que el uso de estas estimaciones más precisas permita mejorar la validez externa con respecto a la validez obtenida con las puntuaciones directas. Se presentan algunos resultados basados en el modelo logístico de dos parámetros, y se discuten tres aspectos: (a) Cómo deben interpretarse los coeficientes de validez basados

en puntuaciones TRI; (b) como pueden hacerse inferencias con respecto a estos coeficientes, y (c) qué mejoras en validez cabe esperar cuando el modelo es correcto y se usan las puntuaciones TRI en lugar de las puntuaciones directas. Cuatro estudios empíricos en personalidad aportan evidencia acerca de los resultados que cabe esperar en aplicaciones reales, en las que el modelo no es correcto sino tan sólo una aproximación. En todos los ejemplos, el resultado general es que los coeficientes de validez basados en puntuaciones TRI son muy similares a los obtenidos con puntuaciones directas.

REFERENCES

Bechger, T.M., Maris, G., Verstralen,H.H.F.M. & Béguin, A.A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*, 319-334.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick. *Statistical theories of mental tests scores* (pp. 397-472). Reading: Addison-Wesley.

Carrigan, P.M. (1960). Extraversion-introversion as a dimension of personality: A reappraisal. *Psychological Bulletin, 57*, 329-360.

Cooke, D.J. (1995). Psychopathic disturbances in the Scottish prison population: The cross-cultural generalisability of the Hare psychopathy checklist. *Psychology, Crime and Law, 2,* 101-108.

Crowne, D.P. & Marlowe, D. (1964). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349-354.

Eysenck, H.J. & Eysenck, S.B.G. (1963). *Manual of the Eysenck personality inventory.* London: University of London Press.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

Ferrando, P.J. (1999). Likert scaling using continuous, censored and graded response models: Effects on criterion-related validity. *Applied Psychological Measurement, 23,* 161-175.

Ferrando, P.J. (2004). Improving the validity of personality scores with item factor analysis. *Metodología de las Ciencias del Comportamiento, 5,* 197-210.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7,* 385-400.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston: Kluwer.

Hancock, G.R. (1997). Correlation/validity coefficients disattenuated for score reliability: A structural equation modeling approach. *Educational and Psychological Measurement, 57,* 598-606.

Hare, R.D. (1970). *Psychopathy: Theory and research.* New York: Wiley.

Hare, R.D. (1991). *The Hare Psychopathy Checklist –Revised. Toronto*: Multi-Health Systems.

Jöreskog, K.G. & Sörbom, D. (1996). *LISREL8: User's Reference Guide [Computer software]*. Chicago, IL: Scientific Software.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp 159-168). Greenwhich, CT: JAI.

Lord, F.M. (1952). A theory of test scores. *Psychometrika Monograph Supplement.* No. 7.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233-245.

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

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251-280.

MacDonald, P. & Pauonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62,* 921-943.

McBride, J. R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp 224-236). New York: Academic Press.

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293-299.

Messick, S. (1993). Validity. In R.L. Linn, (Ed.) *Educational measurement third edition* (pp. 13-103). Phoenix: Oryx Press.

Nunnally, J.C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Reise, S.P. (1999). Personality measurement issues viewed through the eyes of IRT. In S.E. Embretson & S.L. Hershberger (Eds). *The new rules of measurement* (pp. 219-241). Hillsdale,NJ: LEA.

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

Reynolds, C.R. & Richmond, B.O. (1978). What I think and fell: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6,* 271-280.

Rocklin, T. & Revelle, W. (1981). The measurement of extraversion: A comparison of the Eysenck personality inventory and the Eysenck personality questionnaire. *British Journal of Social Psychology, 20*, 279-289.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement. No. 34.*

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika, 42,* 163-191.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229-244.

Schmidt, F.L. & Hunter, J.E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183-198.

Stanley, J.C. & Wang, M.D. (1969). Restrictions of the possible values of r_{12} given r_{13} and r_{23}. *Educational and Psychological Measurement, 29,* 579-581.

Young, J.W. (1995). A comparison of two adjustment methods for improving the prediction of law school grades. *Educational and Psychological Measurement, 55,* 558-571.

Zimowski, M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (2002). *BILOG MG II: Multiple group item analysis and test scoring with binary logistic models [Computer software]*. Chicago, IL: Scientific Software.

(Manuscript received: 25 July 2006; accepted: 20 November 2006)