Psicológica (2012), 33, 111-134.

Assessing the discriminating power of item and test scores in the linear factor-analysis model

Pere J. Ferrando*

Rovira i Virgili University, Spain

Model-based attempts to rigorously study the broad and imprecise concept of 'discriminating power' are scarce, and generally limited to nonlinear models for binary responses. This paper proposes a comprehensive framework for assessing the discriminating power of item and test scores which are analyzed or obtained using Spearman's factor-analytic model. The proposed framework is organized on the basis of three criteria: (a) type of score, (b) range of discrimination, and (c) conceptualization and aspect that are measured. Within this framework, the functioning and interpretation of 16 measures, of which 6 appear to be new, are discussed, and the relations between them are established. The usefulness of the proposal in psychometric FA applications is illustrated by means of an empirical example.

As several authors have pointed out (Loevinger, 1954; Lord & Novick, 1968; McDonald, 1999) the term "discriminating power" is rather imprecise. In a broad sense, it refers to the degree to which a score varies with trait level, as well as the effectiveness of this score to distinguish between respondents with a high trait level and respondents with a low trait level. This property is directly related to the quality of the score as a measure of the trait (Lord & Novick, 1968; McDonald, 1999) so it is of central practical importance, particularly in the context of item selection. For this reason, most research has focused on developing indices that are thought to express this property numerically, whereas more theoretically-oriented research is far scarcer. Below we provide a review of the literature that is most related to the present developments. The review is organized

* The research was partially supported by grants from the Spanish National Organization of the Blind (ONCE), the Catalan Ministry of Universities, Research and the Information Society (2005SGR00017), and the Spanish Ministry of Education and Science (PSI2008-00236/PSIC). Correspondence should be sent to:Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicología. Carretera Valls s/n. 43007 Tarragona (Spain). E-mail: perejoan.ferrando@urv.cat

around three classification criteria: (a) type of score, (b) range of discrimination, and (c) conceptualization, and the aspect that is measured.

As for criterion (a), in principle, the discriminating power can be assessed for all types of score (Lord, 1980; McDonald, 1999). However, most research has focused on direct scores: either single-item scores or (to a far lesser extent) total-test scores obtained by the unweighted sum of the item scores (e.g. Levine & Lord, 1959). The discriminating power of more complex scoring schemas such as maximum likelihood trait estimates has only been indirectly considered for some item response models via the related concept of test information (Lord, 1980; McDonald, 1999).

The term "range of discriminating power" for defining criterion (b) was proposed by Loevinger (1954) and we shall use Mellenbergh's (1996) distinction between population-independent measurement versus population-dependent measurement to discuss it. The criterion refers to whether discriminating power is assessed at a single point or trait level (population independent) or over the entire trait distribution in a given population (population dependent). Standard item discrimination indices used in classical test theory (CTT), such as the upper-lower index or the item-total correlation, assess the item discriminating power in a given population (Lord, 1980). Population-independent measures, both at the item and the total-test level, were developed subsequently, and aimed at more complex models in which measurement precision was assumed to be different at different trait levels. Lawley (1943) and Lord (1952) proposed the first indices of this type. More recently proposed measures of item and test information derived from certain item response theory (IRT) models can be considered as population-independent measures of discrimination (Lord, 1980, sec. 5.2; Nicewander, 1993).

The relatively few studies that have dealt with criterion (c) have tried to arrive at a more precise conception of discrimination by defining more specific aspects of the general property. Loevinger (1954, Loevinger, Glaser & DuBois, 1953) distinguished two aspects that we shall consider here: fineness and probability. Discriminating fineness refers to the magnitude of the trait differences that the score is able to discriminate. It is this aspect that most existing measures of discrimination attempt to measure.

Several authors (Lawley, 1943; Levine & Lord, 1959; Lord, 1952; Mandel & Stiehler, 1954; Nicewander, 1993) further distinguished two components of fineness, and discussed them in terms of population-independent discrimination. The first component refers to the magnitude of the expected score difference for the given trait difference (at the point at which this difference is considered). This component is generally assessed

by the steepness of the score-trait slope at this point. The second component refers to the amount of measurement error of the score at this point. In terms of information theory (see e.g. Cronbach & Glesser, 1964; or Nicewander, 1993) the first component can be interpreted as the strength of the transmission, and the second as the strength of the interference. Now, effective discriminating fineness is associated with both a high expected difference or steep slope, and a small error variance (Levine & Lord, 1959). So, all the authors referred to above proposed to combine both components in a single measure of the signal-to-noise-ratio type (Cronbach & Glesser, 1964; Nicewander, 1993).

Discriminating probability refers to a basic question which arises when assessing 'relations of difference' (Ferguson, 1949): does an observed difference between the scores of two individuals reflect a 'true' difference in their trait levels which goes in the same direction? Operationally, discriminating probability is measured by the proportion of discriminations which are in the same direction as trait differences (Loevinger, 1954; Milholland, 1955). Thurlow (1950), Ferguson (1949) and Milholland (1955) proposed discrimination indices that can be regarded as measures of probability. According to the present criteria, they are all population-dependent measures intended for raw test scores. These indices have seldom been used, although recently they seem to be arousing some interest again (Hankins, 2007), and their main limitation is that they are difficult to link to an specific response model. The view adopted here is that probability indices are mostly useful as auxiliary measures of discrimination that provide additional information, and that enable the results obtained by using fineness measures to be interpreted more clearly.

The review of the literature shows that most model-based research on discrimination has focused on binary items and scores derived from these items. Binary items are still widely used, especially in ability measurement. However, in the attitude and personality domains, it is more common to use more continuous formats (e.g. Dawes, 1972; Ferrando, 2002) such as graded-responses with a sizable number of points, or even more continuous, such as graphic or visual scales. These types of items are generally calibrated using linear factor analysis (FA; e.g. Ferrando, 2002, 2009; Hofstee, Ten Berge & Hendricks, 1998; McDonald, 1999). Now, linear FA is a model for continuous-unlimited variables whereas item responses are bounded and, to a greater or lesser extent, discrete. So, the model must be viewed as an approximation. This simple linear approximation, however, seems to work well in most applications (Atkinson, 1988; Ferrando, 2002, 2009; Hofstee et al., 1998; McDonald, 1999), especially those based on the type of items discussed above.

In spite of its widespread use as an item response model, there seems to be no systematic study on the types of discrimination measures that can be used with the FA model. However, this statement should perhaps be qualified. Some widely used binary and graded-response IRT models can be parameterized as item FA models by using a formulation based on hypothetical underlying response variables (e.g. McDonald, 1999; Reckase, 1997). The relations between both parameterizations, in turn, enable relations between IRT and FA measures of item discrimination to be obtained (see e.g. Kamata & Bauer, 2008, equation 2). As discussed above, however, this paper deals with the linear FA model when it is applied directly to the observed item responses. In particular, we shall focus on assessing the discriminating power of item and test scores which are analyzed or constructed using Spearman's (1904) unidimensional FA model. In the psychometric literature, when Spearman's model is used as an item response model it is also known as the congeneric test (item) score model (Jöreskog, 1971), a name that we shall use here.

The present study provides a general framework which describes the properties of 16 discrimination measures, how they work and how they are related. As far as I know, this is the first comprehensive treatment of the discrimination measures that can be used in psychometric applications of the congeneric model. It mainly aims to be of interest to applied researchers, to serve as a guide to choosing the most appropriate discrimination measures for each particular study, hopefully improving both analyses and interpretation and leading to better applications of the model.

Many of the measures discussed here are already known or can be considered as particular cases of general existing measures. However, the study also provides new theoretical and methodological contributions. At the theoretical level, some interpretations of existing measures, and the relations between them, which do not seem to have been considered to date are discussed. At the methodological level, six of the measures proposed, which are probability measures, appear to be new. Unlike existing measures of this type that cannot be linked to specific theory, the newly developed measures are directly derived from the congeneric model.

Review of the Model, Basic Results, and General Scheme

In this paper the congeneric model is conceptualized as a linear IRT model intended for (approximately) continuous responses (Ferrando, 2009; Mellenbergh, 1994; Thissen, Steinberg, Pyszcynski & Greenberg, 1983). More in detail, we shall adopt the same unified framework as in McDonald (1982, 1999) and Lord and Novick (1968, chapter 24), and consider the

congeneric model as a particular case of a general IRT (or latent trait) model described by the principle of local independence.

Consider a set of n items that measure a single trait θ. The congeneric model for the observed score in item j is:

$$X_j = \mu_j + \lambda_j \theta + \varepsilon_j \tag{1}$$

where X_j is the observed item score, μ_j is the item intercept, λ_j the item loading, slope, or regression weight, and ε_j the measurement error. For fixed θ, the item scores are distributed independently (this is the IRT principle of local independence that characterizes the general model). The conditional distribution is assumed to be normal, with mean and variance given by

$$E(X_j \mid \theta) = \mu_j + \lambda_j \theta \quad ; \quad Var(X_j \mid \theta) = \sigma^2_{\varepsilon j}. \tag{2}$$

The two expressions in (2) summarize the main differences between the congeneric model and most standard IRT models intended for discrete responses. These differences are particularly clear with respect to the two-parameter model (2PM) for binary responses. The conditional mean in the first expression in (2) is, by definition, the item characteristic curve (ICC) of the congeneric model (see e.g. Ferrando, 2009), which is indeed a straight line. The conditional variance in the second expression is constant and does not depend on the trait level (i.e. homoscedasticity). In contrast, in the 2PM the ICC is a sigmoid curve (an ogive), and the conditional variance is generally different at different trait levels (see e.g. McDonald, 1982, 1999). As discussed below, these differences will be relevant regarding the measures discussed in this paper but, even so, we give θ the same meaning as it has in the standard IRT models. It is the quantitative attribute or characteristic of the respondents that the set of items measure in common. What is different in linear and nonlinear models, then, is the link function that relates the θ levels to the observed item scores.

In model (1) the item scores X_j are observed, and so the response variables are scaled (the scale depending on the scoring schema). However, a scaling choice must be made for θ. Kamata and Bauer (2008) discussed the two most common scaling conventions. Either θ is standardized, or the intercept and loading of an item, which acts as an indicator, are fixed. The second scaling allows the mean and variance of θ to be freely estimated. As discussed below, the possibility of θ having different variance in different

populations is a basic issue in the distinction between population-dependent and population-independent measures.

The mean and variance of the unconditional (marginal) distribution of X_j over the entire population are

$$E(X_j) = \mu_j + \lambda_j E(\theta) \quad ; \quad Var(X_j) = \lambda^2{}_j Var(\theta) + \sigma^2_{ej} \tag{3}$$

Consider now two trait levels—θ_1 and θ_2—and let $\delta = \theta_1 - \theta_2$ be the difference between them, which we shall refer to as the 'true' difference. Let $d_j = X_{j1}-X_{j2}$ be the corresponding difference in the observed scores in item j. For fixed δ, and according to the model assumptions, the conditional distribution of the observed differences is normal, with mean and variance given by

$$E(d_j \mid \delta) = \lambda_j \delta \quad ; \quad Var(d_j \mid \delta) = 2\sigma^2_{ej} \tag{4}$$

In most applications, estimation of model (1) has two stages. In the first stage (item calibration), the item parameters (intercepts, loadings and error variances) are estimated and the global fit of the model is assessed. In the second stage (scoring), provided that the fit is acceptable, the item estimates are taken as fixed and known values, and individual trait estimates or scores based on the entire set of items are obtained. In this paper we shall assess the discriminating power of two types of total-test scores (see McDonald, 1999): raw scores obtained as the simple sum of the individual item scores, and maximum likelihood estimates (MLEs).

The results that we need for the raw scores can be derived directly from the item results discussed above. Let $X=\Sigma X_j$ be the sum of the n item scores. From the normality and local independence assumptions it follows that, for fixed θ, X is the sum of n independent normal variables. Now let δ be as defined above, and $d = X_1-X_2$ be the corresponding difference in the observed raw scores. For fixed δ, the conditional distribution of the observed differences is normal, with mean and variance given by

$$E(d \mid \delta) = \delta \sum_j^n \lambda_j \quad ; \quad Var(d \mid \delta) = 2\sum_j^n \sigma^2_{ej} \ . \tag{5}$$

We turn now to the MLE, which we shall denote by $\hat{\theta}$. Under the assumption of conditional normality, the MLE for respondent *i* is the Bartlett estimated factor score (McDonald, 1982; Mellenbergh, 1994). Asymptotically (as the number of items increases without limit) the conditional distribution of $\hat{\theta}$ for fixed θ is normal with mean and variance given by

$$E(\hat{\theta} \mid \theta) = \theta \quad ; \quad Var(\hat{\theta}_i \mid \theta_i) = (I)^{-1} \tag{6}$$

where the conditional variance $Var(\hat{\theta}_i \mid \theta_i)$ is the variance of the estimate error, and *I* is the test information function, defined as

$$I = \sum_j^n I_j \quad ; \quad I_j = \frac{\lambda_j^{\,2}}{\sigma_{\varepsilon j}^2}. \tag{7}$$

The term I_j is item *j*'s contribution to the test information or, more simply, the item information (Mellenbergh, 1994). Now, again let δ be as defined above, and $d(\hat{\theta})$ be the corresponding difference in the MLEs. For fixed δ, the conditional distribution of the $d(\hat{\theta})$'s is (asymptotically) normal, with mean and variance given by

$$E(d(\hat{\theta}) \mid \delta) = \delta \quad ; \quad Var(d(\hat{\theta}) \mid \delta) = 2(I)^{-1}. \tag{8}$$

Equation (8) completes the results that are needed to interpret and or develop the measures of discrimination that we shall discuss in the following sections. To make the discussion clearer, table 1 shows all the measures categorized according to the type of score, range of discrimination, and conceptualization/aspect. For each measure, table 1 provides the mathematical formula and the number of the equation in which it is defined in the article given in brackets.

Table 1. Measures of discriminating power for the congeneric model.

	Population-independent				Population-dependent			
	Fineness		Probability		Fineness		Probability	
Item Score	λ_j	(1)	$\Delta c_j = \Phi(Dc_j)$	(12)	$SC_j = Dc_j\sqrt{Var(\delta)}$	(13)	$\Delta u_j = \sum_{k=1}^{q}\Phi(Dc_j X_k)W(X_k)$	(16)
	$Dc_j = \dfrac{\lambda_j}{\sqrt{2\sigma^2_{ej}}}$	(9)			$\alpha_j = \sqrt{\dfrac{1}{1+(SC^2{}_j)^{-1}}}$	(15)		
Raw Score	$Dc = \dfrac{\sum_j^n \lambda_j}{\sqrt{2\sum_j^n \sigma^2_{ej}}}$	(17)	$\Delta c = \Phi(Dc)$	(19)	$SC = \sqrt{D^2 c Var(\delta)}$	(20)	$\Delta u = \sum_{k=1}^{q}\Phi(Dc X_k)W(X_k)$	(23)
					$\alpha = \sqrt{\dfrac{1}{1+(SC^2)^{-1}}}$	(22)		
ML Score	$Dc(\hat{\theta}) = \dfrac{1}{\sqrt{2(I)^{-1}}}$	(18)	$\Delta c(\hat{\theta}) = \Phi(Dc(\hat{\theta})$	(19)	$SC(\hat{\theta}) = \sqrt{D^2 c(\hat{\theta}) Var(\delta)}$	(20)	$\Delta u(\hat{\theta}) = \sum_{k=1}^{q}\Phi(Dc(\hat{\theta})X_k)W(X_k)$	(23)
					$\alpha(\hat{\theta}) = \sqrt{\dfrac{1}{1+(SC(\hat{\theta})^2)^{-1}}}$	(22)		

Measures of Item Discriminating Power

Population-independent measures

In previous proposals of model (1), the weight λ_j was considered as the basic item discrimination index (Ferrando, 2002, 2009; McDonald, 1999; Mellenbergh, 1994; Thissen et al., 1983). From equation (4) it follows that λ_j can be interpreted as the expected difference in item j's score corresponding to a 'true' unit difference in θ (i.e. δ=1). So, λ_j is a population-independent measure of fineness: it gives the expected score difference for fixed trait difference (i.e. at a single point) regardless of the trait distribution and variance.

It might be instructive to compare λ_j as a measure of fineness in the linear model to the item slope parameter (item discrimination) in the 2PM, where the ICC is an ogive, so the slope of the curve varies at different trait levels. To solve this indeterminacy, the discrimination parameter is defined as (proportional to) the slope at the point at which it is maximal (the inflexion point of the ICC). So, the item discrimination parameter in the 2PM is a population-independent measure of fineness that assesses the ability of the item to discriminate at a specific trait level (the point of

maximal discrimination; see e.g. Reckase, 1997). In contrast, in the congeneric model the slope λ_j is a constant that does not depend on θ. So, λ_j is not specific but general: the fineness is the same at any point on the trait continuum.

As a measure of discriminating fineness, λ_j assesses only one of the two components discussed above: the strength of the transmission. Furthermore, its values depend on the measurement scale of the item scores as well as on the scaling choice for θ. So, λ_j is an incomplete measure whose values are difficult to interpret. In our view, λ_j is a basic item parameter that should be reported and considered. Ding and Hershberger (2002), for example, suggested testing λ_j for significance as the first step in deciding whether the item can be considered as a measure of the trait at all. If it can, λ_j is clearly insufficient as a single measure of item discriminating power.

A more complete measure of the signal-to-noise-ratio type can be derived from the results in equation (4). The proposed measure is

$$Dc_j = \frac{\lambda_j}{\sqrt{2\sigma^2_{\varepsilon j}}} \tag{9}$$

The index *Dcj* is a population-independent measure of fineness that assesses both components: strength of the transmission/signal in the numerator, and strength of the interference/noise in the denominator. It can be interpreted as the ratio between the expected score difference discussed above and the amount of overlap in the conditional distributions of the X_{j1} and X_{j2} values.

In real applications λ_j is assumed to be always positive. So *Dcj* is bounded below by zero and has no upper bound. It does not depend on the measurement scale of X_j so it can be interpreted by setting standards of magnitude (possibly based on empirical evidence). Furthermore, interpretation of *Dcj* can be enhanced if it is viewed as the expected value of a critical ratio. In effect, for a 'true' unit difference, the probability that the corresponding observed difference is considered as statistically significant is given by

$$p(Dc_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\lambda_j - z_c\sqrt{2\sigma^2_{ej}}}{\sqrt{2\sigma^2_{ej}}}} \exp(-\frac{1}{2}u^2)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Dc_j - z_c} \exp(-\frac{1}{2}u^2)du = \Phi(Dc_j - z_c). \quad (10)$$

where z_c is the critical value associated to the chosen significance level. Thus $p(Dcj)$ can be interpreted as the probability that a 'true' unit difference be detected by item j's scores. Result (10) follows from the assumption that the conditional distribution of d_j is normal with the mean and variance given in (4)

Although strictly speaking *Dcj* appears to be a new proposal based on conditional distribution (4), it is directly related to already existing measures that were proposed either with no reference to a specific model or with reference to a different model. Thus, the λ-ratio proposed by Lawley (1943) for the 2-parameter normal-ogive model becomes *Dcj* when developed for the congeneric model. When the general discrimination index proposed by Lord, (1952), Levine & Lord (1959) and Mandel & Stiehler (1954) is applied to the congeneric model, it becomes proportional to *Dcj*. Finally, the square of *Dcj* is proportional to the item information (see equation 7). The interpretation of *Dcj* as an expected critical ratio and result (10) appear to be new.

We turn now to the population-independent measures of discriminating probability. For a fixed 'true' difference δ (in absolute value), the probability that the expected difference d_j is in the same direction as the 'true' difference is found to be

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\lambda_j}{\sqrt{2\sigma^2_{ej}}}\delta} \exp(-\frac{1}{2}u^2)du = \Phi\left(\frac{\lambda_j}{\sqrt{2\sigma^2_{ej}}}\delta\right) = \Phi(Dc_j\delta). \quad (11)$$

By using the same rationale as above we propose as a measure of probability the p value in (11), which corresponds to a 'true' unit difference (i.e. δ=1).

$$\Delta c_j = \Phi(Dc_j). \quad (12)$$

Mathematically, Dc_j is the normal equivalent deviate or the probit transform of Δc_j. Conceptually, Δc_j can be interpreted as the probability that the observed difference in item *j*'s score is in the same direction as the 'true' difference in the trait level when this 'true' difference has a value of 1. Given that Dc_j and Δc_j are transformations of each other, the main interest of Δc_j is as an auxiliary measure which provides additional information. For example, assume that Dc_j=0.50 and Dc_k=1.50. In item *j*'s scores, the signal has only half the strength of the noise, while for item *k* the signal is 1.5 times stronger than the noise. So, item *k* has more discriminating fineness than item *j*. The corresponding probability values are Δc_j=0.69 and Δc_k=0.93. So, for a 'true' unit difference, the probability that item *j*'s score reflects a difference in the same direction is only 69%, whereas the probability for item *k*'s score is 93%.

Measures (9) and (12) are more complete and informative than λ_j and their values do not depend on the measurement scale of the item scores. However, they do depend on the scaling choice which is made for θ, and this fact must be considered when interpreting the results they provide. As one reviewer noted, a 'true' unit difference has a different meaning depending on the scale on which the differences are measured. Further discussion on this point is provided in the measures proposed below.

Population-dependent measures

A plausible proposal for a population-dependent measure of fineness (e.g. Jackson, 1939) is to combine the conditional amount of finesses at a given 'true' difference with the magnitude of the true differences in the population. So, if an item score has a high amount of fineness for discriminating at a given difference level, but the differences in the population are consistently small (i.e. the trait levels are all similar), the fineness in the population will be low. On the other hand, if the population-independent fineness is only moderate but the population trait levels are very different, the overall fineness in the population will be high.

The sensitivity coefficient (SC) proposed by Jackson (1939) can be shown to be a direct combination of both components. In the congeneric model it is given by

$$SC_j = \sqrt{\frac{\lambda_j^2 Var(\theta)}{\sigma_{\varepsilon j}^2}} = Dc_j \sqrt{Var(\delta)}\,. \qquad (13)$$

The relation between both expressions in (13) is obtained from the result $Var(\delta)=2Var(\theta)$, which is discussed below in detail (see discussion regarding equation 15). In the context of CTT in which was proposed, SC_j is the square root of the ratio between 'true' variance and error variance. So, it is bounded below by zero and has no upper bound. Here, we interpret it as the amount of fineness item *j* has for discriminating a 'true' unit difference (i.e. Dc_j) weighted by the typical difference in the population of interest (i.e. $\sqrt{Var(\delta)}$, and note that $\sqrt{Var(\delta)}$ is also the root mean square of the differences).

In applications of the FA model, however, the population-dependent measure of discrimination that is routinely used is the standardized item weight, denoted here by α_j. This is, indeed, the weight obtained when the inter-item correlation matrix is factor analyzed. It is given by

$$\alpha_j = \lambda_j \sqrt{\frac{Var(\theta)}{Var(X_j)}} = \lambda_j \sqrt{\frac{Var(\theta)}{\lambda_j^2 Var(\theta) + \sigma_{ej}^2}} \,. \tag{14}$$

As a measure of discrimination, α_j has important advantages. Under the assumptions considered here, it is a normed index, bounded between 0 and 1, which can be interpreted as the product-moment correlation between θ and the item score. So, it is a "classical" measure of fineness with the same rationale as the item-total correlation index (Henrysson, 1962). Furthermore, in the context of CTT, α_j is the square root of the ratio between 'true' variance and total variance. So, its squared value can be interpreted as the item reliability coefficient (Lord & Novick, 1968, sec. 3.4). As far as the present considerations are concerned, however, it is of more interest to assess the relation between α_j and SC_j. The relation is

$$\alpha_j = \sqrt{\frac{1}{1 + \left(SC^2{}_j\right)^{-1}}} \,. \tag{15}$$

So, α_j can be viewed as a transformation of the sensitivity coefficient that maps its values on the 0-1 interval. In spite of the interpretative advantages of α_j just discussed, it is not clear that it is always preferable to SC_j. Given that SC_j has no upper bound, it follows that when the α_j values are near to one, small changes in α_j lead to large changes in SC_j. So, in these

cases SC_j will be more sensitive for detecting differences or changes in fineness than α_j. At the same time, however, a ratio that increases without bound can become very unstable.

The natural extension of the Δc_j probability measure to the entire population is the proportion of observed differences in the population that go in the same direction as the corresponding 'true' differences, a measure considered by Milholland (1955). To obtain this marginal measure, however, the distribution of δ must first be specified.

In applications of model (1) the distribution that is specified is that of θ. Now, the distribution of δ can be viewed as the distribution of the 'true' differences between all pairs of trait levels in the population (see Milholland, 1955). If so, it follows that the distribution of δ is the same as that of θ, but with zero mean and $Var(\delta)=2Var(\theta)$. (This last result has been used in equation 13).

Once the distribution of δ has been established, it can be approximated as accurately as required using numerical quadrature. A simple procedure is to use rectangular quadrature over q equally spaced points. The probability measure Δu_j is then computed as (see the last term on the right hand side in equation 11)

$$\Delta u_j = \sum_{k=1}^{q} \Phi(Dc_j X_k) W(X_k) \tag{16}$$

where X_k are the nodes and $W(X_k)$ are the weights that approximate the distribution of δ.

Measures of Discriminating Power for Total-Test Scores

For both raw scores and MLEs, the extension of all the indices that were proposed at the item level is rather direct, and can be obtained by using equations (5) and (6), respectively. We shall discuss the score-based extensions in the same order in which they were discussed at the item level.

For the raw scores, the population-independent signal-to-noise measure of fineness Dc, which is the extension of Dc_j, is given by

$$Dc = \frac{\sum_{j}^{n} \lambda_j}{\sqrt{2\sum_{j}^{n} \sigma_{\varepsilon j}^2}} \cdot . \tag{17}$$

And for the MLEs it is given by

$$Dc(\hat{\theta}) = \frac{1}{\sqrt{2(I)^{-1}}} \cdot . \tag{18}$$

Both measures have the same interpretation as in the original item. The Dc and $Dc(\hat{\theta})$ indices measure the amount of fineness of the corresponding score (raw or MLE) in discriminating between two respondents whose 'true' trait levels differ by one unit.

The ratio $Dc(\hat{\theta})/Dc$ is a particular application of the sensitivity ratio (SR) proposed by Mandel and Stiehler (1954). Furthermore, the square of this ratio is the relative efficiency of the MLEs with respect to the raw scores (e.g. McDonald, 1999). It can be shown that the relative efficiency (and, therefore, the SR) is always greater than one except when the Dc_j values are the same for all of the items, in which case it is exactly one (McDonald, 1999). So, the amount of conditional discriminating fineness of the MLEs is always greater than that of the raw scores. In applications, the SR can be used to determine whether the improvement in fineness obtained by using the more complex MLEs is negligible or not.

The auxiliary probability measures that correspond to Dc and $Dc(\hat{\theta})$ are given by

$$\Delta c = \Phi(Dc) \quad ; \quad \Delta c(\hat{\theta}) = \Phi(Dc(\hat{\theta})). \tag{19}$$

And they can be interpreted as the probability that the observed difference in the corresponding score (raw or MLE) is in the same direction as the 'true' trait difference when this 'true' difference is one.

We turn now to the population-dependent measures. The direct extensions of the item sensitivity coefficient to the raw and MLE scores are given by

$$SC = Dc\sqrt{Var(\delta)} \quad ; \quad SC(\hat{\theta}) = Dc\sqrt{(\hat{\theta})Var(\delta)} \quad . \tag{20}$$

In the CTT context, the square of the SC on the left-hand-side of (20) has a one-to-one relation with the well known omega (ω) reliability coefficient (e.g. McDonald, 1999). SC-squared is a ratio of true variance to error variance while omega is the ratio of true variance to total variance. So, the relation is

$$SC^2 = \frac{\omega}{1-\omega}. \tag{21}$$

According to the present framework, and, for the corresponding score, the two indices on (20) reflect the two components discussed above: (a) the amount of fineness at a unit difference level, and (b) the typical difference in the population.

The corresponding α transformations are

$$\alpha = \sqrt{\frac{1}{1+(SC^2)^{-1}}} \quad ; \quad \alpha(\hat{\theta}) = \sqrt{\frac{1}{1+(SC(\hat{\theta})^2)^{-1}}} \; . \tag{22}$$

As in the item case, both α and $\alpha(\hat{\theta})$ can also be interpreted in ways that are more related to CTT-based analysis. Each index can be interpreted as the product-moment correlation between the trait and the corresponding score. Its square can be interpreted as the reliability of the corresponding score (raw or MLE; e.g. Mellenbergh, 1994).

Finally, the population-dependent probability measures, which are the extensions of *Δuj*, can be obtained as

$$\Delta u = \sum_{k=1}^{q} \Phi(DcX_k) W(X_k) \quad ; \quad \Delta u(\hat{\theta}) = \sum_{k=1}^{q} \Phi(Dc(\hat{\theta})X_k) W(X_k). \tag{23}$$

As in the item case, and for each type of score, the index Δu is interpreted as the proportion of observed differences in the population that are in the same direction as the corresponding 'true' differences.

ILLUSTRATIVE EXAMPLE

To illustrate the measures discussed in this article we shall use a multiple-group application of the congeneric model. This type of application was chosen to show how, under certain item invariance conditions, some measures of discrimination do not depend on the group (population) in which they are assessed, whereas other do. A short 5-item scale for measuring Extraversion was administered to two groups of undergraduate students. The items used a 5-point Likert format, and, for interpretative purposes, the scores were scaled in the 0-1 interval. In group 1 (N=455), participants were asked to respond to the scale under standard instructions, which, among other things, advised them to give honest answers. In group 2 (N=418) participants were instructed to imagine they were applying for a job and to try to give a good impression regardless of the truthful answer.

Because the model is small and the samples are reasonable large, the congeneric model in equation (1) was fitted using full weighted least squares estimation as implemented in the Mplus program version 5 (Muthén and Muthén, 2007). The item weights and error variances in equations (1) and (2) were constrained to be equal in both groups (see Millsap & Meredith, 2007). Conceptually the restriction implies that each of the items has the same quality as a measure of the trait under honest and faking-good conditions. The two scaling conventions discussed by Kamata and Bauer (2008) were used for fitting the model. In group 1 θ was standardized. In group 2, and given the invariance constraints just discussed, the fixed values of the item parameters allowed the variance of θ to be freely estimated in this group (relative to the fixed unit variance in group 1). Model-data fit was reasonably good. The chi-squared test statistic value with 19 degrees of freedom was $\chi^2(19)=39.12$, the root mean squared error of approximation (RMSEA) and its 95% confidence interval were 0.049 and (0.025; 0.072), respectively, and the Non-Normed Fit index was 0.97.

The variance estimates of the trait were 1.00 in group 1 (fixed) and 0.30 in group 2. So the corresponding δ variances were 2.00 and 0.60. The reduced variance in group 2, then, suggests that under faking-motivating instructions participants tend to respond in a more similar way, thus reducing the inter-individual differences in this respect. As discussed above,

this homogenization of the group will decrease the population-dependent discriminating power in all types of scores.

The results on all the measures of discrimination discussed in this paper are shown in table 2.

Table 2. Assessment of discriminating power. Illustrative example

	Population-independent			Population-dependent					
				SC_j		$.\alpha_j$		Δu_j	
	λ_j	Dc_j	Δc_j	G1	G2	G1	G2	G1	G2
i1	0.15 (0.01)	0.61	0.73	0.86	0.47	0.66	0.43	0.73	0.64
i2	0.21(0.01)	1.04	0.85	1.47	0.80	0.83	0.62	0.82	0.72
i3	0.14(0.01)	0.49	0.69	0.69	0.38	0.57	0.35	0.70	0.62
i4	0.16(0.01)	0.48	0.68	0.67	0.37	0.56	0.35	0.69	0.61
i5	0.13(0.01)	0.67	0.75	0.95	0.52	0.68	0.45	0.74	0.65
				SC		α		Δu	
		Dc	Δc	G1	G2	G1	G2	G1	G2
Raw		1.38	0.92	1.95	1.06	0.89	0.73	0.86	0.77
MLE		1.54	0.94	2.17	1.19	0.91	0.76	0.88	0.79

Before interpreting the specific results, we first note that, given the invariance constraints which were imposed, the population-independent measures of discrimination are the same in both groups, and so they are reported only once with no group reference. As discussed above, they measure discrimination for a fixed trait difference, regardless of the trait distribution and variance. Second, we note that the standard errors of the λ's are very small, which is to be expected given the reasonable sample sizes and the simplicity of the model.

We first discuss the population-independent item measures. If the λ values are compared to their respective standard errors, it is clear that the discriminating power of all the items can be considered as significantly different from zero. As for interpretation, however, the λ values only indicate that a unit change in the trait level results in an expected change between 0.13 and 0.21 units in the 0-1 item score scale, which is insufficient for evaluating the amount of discriminating fineness of the item scores. Note that, except for item 2, the λ values are quite similar for the remaining items, a typical result in personality items (Ferrando, 2002, 2009).

The Dc_j values provide more useful information. First, because both components of fineness are now taken into account, the differences between the items are accentuated, which allows for finer comparisons. As for interpretations, in all of the items except item 2, the signal is smaller than the noise. For the worst items (3 and 4) the expected difference for unit δ is less than half of the standard error of the differences. If the Dc_j values are interpreted as expected values of a critical ratio, it is clear that they are far smaller than the cut-off value for considering significance at the 0.05 level (z_c=1.65). So, the probability that a 'true' unit difference might lead to a difference being detected as significant at the 0.05 level is only 0.12 (see equation 10). For the best item (2) the expected difference is slightly larger than one standard error, the expected value is still below the cut-off value, and $p(1.04)$=0.27.

The auxiliary Δc_j values help to complete the interpretation. For unit δ the expected percentage of differences that are in the same direction as the 'true' difference would be between 0.68 and 0.85. In other words, for the worst items 32% of the differences are expected to have a sign that is opposite the sign of the 'true' difference. Overall, it seems clear that the discriminating fineness of these items is rather modest, a result that is common to most personality items (Ferrando, 2002, 2009).

We turn now to the population-independent item measures. In group 1, the typical 'true' difference is $\sqrt{2} = 1.41$, larger than 1. So the SCj's are larger than the corresponding Dc_j's. Conceptually this means that the amount of item fineness at a unit difference is somewhat improved in a population in which differences larger than 1 are common. The corresponding α values, if interpreted as item-trait correlations, appear to be acceptable. However, the marginal proportion of differences which are expected to be in the same direction as 'true' differences is still quite low, even for the best item. These proportions were assessed by assuming that the distribution of δ was normal. Overall, the increase in discriminating fineness in this group is not accompanied by a clear increase in discriminant probability.

As expected, the situation is far worse in group 2. Here, the typical difference is $\sqrt{0.6} = 0.77$. If the power for discriminating a unit trait difference is poor, the marginal power in a population in which most differences are lower than 1 would be expected to be very low, and this appears to be the case according to all the measures.

We turn now to the total-test scores. As for the population-independent measures, in spite of the reduced number of items, it appears that the discriminating power of the test scores is considerably better than

that of the individual items, and the signal is here clearly stronger than the noise. If the *Dc*'s are interpreted as expected critical-ratio values, both values are still below the 1.65 cut-off point, but are quite close. For the raw scores the probability that a 'true' unit difference might lead to a difference being detected as significant at the 0.05 level is: $p(1.38)= 0.40$. For the MLEs it is: $p(1.54)= 0.46$. As auxiliary measures, the proportions of differences expected to go in the same direction as the 'true' unit difference are far better than in the individual-item case: 92% and 94%. Finally, as for the improvements of the MLEs over the raw scores, the sensitivity ratio is 1.12. Given that the discriminating power of the test scores is still quite modest it seems reasonable to take advantage of the small gains provided by the MLEs.

The differential amount of population-dependent discrimination in both groups at the item level are also reflected in the total scores although with improvements in all cases. And these improvements allow the differences between some indices to be better appreciated. Thus, in group 1, in which the α values are rather high, the difference between raw scores and MLEs in terms of α is small, but it becomes clearer in terms of SC_j. Overall, in the best case, the MLEs in group 1 would have an acceptable (for a short personality test) amount of discriminating power both in terms of fineness and probability. On the other hand the discriminating power would still be poor in group 2.

DISCUSSION

Most psychometric applications based on (approximately) continuous items use linear FA as the response model for calibrating items and scoring individuals. In spite of the generalized use and relative simplicity of the model, however, the procedures that are routinely used, as well as the interpretation of results are generally very improvable (Reckase, 1997). As for the issues discussed here, the most general problems seem to be of two types. First, most applications use less information than is provided by the data. Second, the choice of the discrimination measures that are reported and interpreted is guided more by the type of data used as input than by the objectives of the analysis. For example, whether unstandardized (λ's) or standardized (α's) weights are used depends more on whether the input matrix is covariance or correlation than whether population-independent or population-dependent discrimination is of most interest. Overall, and as discussed above, the present study is expected to be useful for improving these two problems.

We shall start with the first problem. The present study discusses measures of discriminating power both at the item and at the total score level. At the item level, the most immediate application is in the process of item analysis. The present proposals allow the researcher to make a detailed assessment of the item discriminating power that is far more complete than the one that is habitually used in applications. In this way, the basic λ_j index is compared to its standard error and its significance is assessed. If it is found to be significant, however, the amount of population-independent item discriminating fineness can be assessed much more completely by using $Dc_{j.}$ and its auxiliary measure Δc_j. The Dc_j measure combines the two aspects of fineness, does not depend on the item scale and can be clearly interpreted both as a signal-to-noise ratio and as the expectation of a critical ratio. As for the population-dependent measures, α_j is a well-known, routinely-used measure that has several meaningful interpretations and is 0-1 normed. However in situations in which the items have very high α_j values, SC_j can be a better measure for comparison purposes. In both cases, the probability measures add non-redundant and useful auxiliary information. As the example shows, the probability values can still be quite low in situations in which the fineness seems to be acceptable. Overall, the results of this detailed scrutiny can be used to select the most discriminating items or for selecting a subset of items that retains the maximum discriminating power when a reduced version of a test is designed.

As discussed above, the discriminating power of test scores is almost never assessed in FA applications. However, the present proposals allow two important points in practical research to be assessed. First, whether the discriminating power is sufficient for the purpose of the test. Second, whether more complex MLEs lead to a substantial increase in the score's discriminating power with respect to that obtained with the simple raw scores. As for the first point, for example, a selection study would mainly require population-independent measures of both fineness and probability while population-dependent measures seem more useful in validity studies carried out in a specific population or in different populations. In both situations, and as in the item case, the present proposal regards fineness measures as the main indicators of discrimination and probability measures as useful auxiliary measures. However, this might also depend on the purposes of the research. Lord (1952), for example, considers a situation in which a given proportion of respondents have been selected on the basis of their test scores, and asks what proportion of the selected individuals really belong to the group with the highest trait levels. This question is best assessed by using a measure of discriminating probability.

We turn now to the second problem. Perhaps the confusion observed in some applications regarding population-independent and population-dependent measures of discrimination partly arises because the distinction is not the same as that in the most common IRT models. In non-linear IRT models the variance of the error of estimate is generally different at different ranges of the trait continuum. So, population-independent measures are generally used for assessing discrimination in a given range, while population-dependent measures are used for assessing the average discriminating power over the whole trait range (e.g. Nicewander, 1993). In the congeneric model, however, the variance of the error of estimate for all the scores considered here remains constant at all trait levels (see e.g. equations 6 and 7). So, in this case, population-independent measures are most appropriate when interest focuses on the discriminating power of the score in general, with no reference to a specific population. Population-dependent measures are most useful when the discriminating power in a specific population is of interest or when possible changes in discrimination over different populations need to be assessed. As the illustrative example shows, population-dependent measures appear to be particularly useful in multiple-group studies. In this type of studies it seems recommendable to follow the strategy of the illustrative examples: first, population-independent measures should be used to get a general idea of the discriminating power of the scores, and then population-dependent measures should be used to assess how the discriminating power changes as a function of the homogeneity/heterogeneity of the group.

We shall finally discuss some limitations of the present study. To start with, it is perhaps better to view the study as an initial, comprehensive proposal that highlights new (or not well known) relations among potential measures of discrimination, and discusses how they can be interpreted. Now, to determine which of these measures will be considered really relevant in future applications requires, indeed, further intensive research, both empirical or simulated. For example, the results of the study suggest that the probability measures add non-redundant and useful auxiliary information, and also that the use of the more complex scoring schemas leads to increases in discriminating power with respect to the simple raw scores. However, results from a single study are hardly generalizable, and these results might be valid only under specific conditions.

The present proposal is limited to the unidimensional FA model. In the domains that most use the items for which FA behaves reasonably well (mainly personality and attitude), single-trait measures are common, but so are multi-trait questionnaires. In the case of FA solutions that approach an independent-cluster structure (see McDonald, 2000), with factorially simple

items, the use of the proposed measures on a separate-scale basis can act as a reasonable approximation. For more factorially complex solutions, it may be interesting to extend the measures proposed here to the multidimensional case. This is left for future research.

As for the feasibility of the proposal, it should be stressed that all the measures considered in this article can easily be computed from a standard FA output by using a calculator or a spreadsheet. However, in the near future the author expects to develop a user-friendly program.

RESUMEN

Evaluación de la capacidad discriminitava de las puntuaciones de los ítems y del test en el modelo de análisis factorial lineal. Las propuestas rigurosas y basadas en un modelo psicométrico para estudiar el impreciso concepto de "capacidad discriminativa" son escasas y generalmente limitadas a los modelos no-lineales para items binarios. En este artículo se propone un marco general para evaluar la capacidad discriminativa de las puntuaciones en ítems y tests que son calibrados mediante el modelo de un factor común. La propuesta se organiza en torno a tres criterios: (a) tipo de puntuación, (b) rango de discriminación y (c) aspecto específico que se evalúa. Dentro del marco propuesto: (a) se discuten las relaciones entre 16 medidas, de las cuales 6 parecen ser nuevas, y (b) se estudian las relaciones entre ellas. La utilidad de la propuesta en las aplicaciones psicométricas que usan el modelo factorial se ilustra mediante un ejemplo empírico.

REFERENCES

Atkinson, L. (1988). The measurement-statistics controversy: Factor analysis and subinterval data. *Bulletin of the Psychonomic Society*, 26, 361-364.

Cronbach, L.J. & Glesser, G.C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement,* 24, 467-480.

Dawes, R.M. (1972). *Fundamentals of attitude measurement. N*ew York: Wiley.

Ding, C.S. & Hershberger, S.L. (2002). Assessing content validity and content equivalence using structural equation modeling. *Structural Equation Modeling,* 9, 283-297.

Ferguson, G.A. (1949). On the theory of test discrimination. *Psychometrika*, 14, 61-68.

Ferrando, P.J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research,* 37, 521-542.

Ferrando, P.J. (2009). Difficulty, discrimination and information indices in the linear factor-analytic model for continuous responses. *Applied Psychological Measurement,* 33, 9-24.

Hankins, M. (2007). Questionnaire discrimination: (re)-introducing coefficient delta. *BMC Medical Research Methodology,* 7, 19.

Henrysson, S. (1962). The relation between factor loadings and biserial correlations in item analysis. *Psychometrika.* 27, 419-424.

Hofstee, W.K.B., Ten Berge, J.M.F., & Hendricks, A.A.J. (1998). How to score questionnaires. *Personality and Individual Differences,* 25, 897-910.

Jackson, R.W.B. (1939). Reliability of mental tests. *British Journal of Psychology,* 29, 267-287.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika,* 36, 109-133.

Kamata, A. & Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling,* 15, 136-153.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh.* 61, 273-287.

Levine, R. & Lord, F.M. (1959). An index of the discriminating power of a test at different parts of the score range. *Educational and Psychological Measurement,* 19, 497-503.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin,* 51, 493-504.

Loevinger, J., Gleser, C.J. & Dubois, P.H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika,* 18, 309-317.

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

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

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

Mandel, J. & Stiehler, R.D. (1954). Sensitivity-A criterion for the comparison of methods of test. *Journal of Research of the National Bureau of Standards,* 53, 155-159.

McDonald, R.P. (1982). Linear vs. non linear models in Item Response Theory. *Applied Psychological Measurement.* 6, 379-396.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah (NJ): LEA.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement,* 24, 99-114.

Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research,* 29, 223-237.

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

Milholland, J.E. (1955). The reliability of test discriminations. *Educational and Psychological Measurement,* 15, 362-375.

Millsap, R.E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R.C. MacCallum (Eds.) *Factor analysis at 100* (pp. 131-152). Mahwah: LEA.

Muthén, L.K. & Muthén, B. (2007). *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Nicewander, W.A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika,* 58, 139-141.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (eds.) *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Spearman, Ch. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology,* 15, 201-293.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An Item Response Theory for personality and attitude scales: using restricted factor analysis. *Applied Psychological Measurement,* 7, 2, 211-226.

Thurlow, W.R. (1950). Direct measures of discrimination among individuals performed by psychological tests. *Journal of Psychology,* 29, 281-314.

(Manuscript received: 10 December 2010; accepted: 21 March 2011)