The Spanish Journal of Psychology
2010, Vol. 13 No. 1, 494-502

ISSN 1138-7416

A General Equation to Obtain Multiple Cut-off Scores on a Test from Multinomial Logistic Regression

Rosa Bersabé and Teresa Rivas
Universidad de Málaga (Spain)

The authors derive a general equation to compute multiple cut-offs on a total test score in order to classify individuals into more than two ordinal categories. The equation is derived from the multinomial logistic regression (MLR) model, which is an extension of the binary logistic regression (BLR) model to accommodate polytomous outcome variables. From this analytical procedure, cut-off scores are established at the test score (the predictor variable) at which an individual is as likely to be in category j as in category j+1 of an ordinal outcome variable. The application of the complete procedure is illustrated by an example with data from an actual study on eating disorders. In this example, two cut-off scores on the Eating Attitudes Test (EAT-26) scores are obtained in order to classify individuals into three ordinal categories: asymptomatic, symptomatic and eating disorder. Diagnoses were made from the responses to a self-report (Q-EDD) that operationalises DSM-IV criteria for eating disorders. Alternatives to the MLR model to set multiple cut-off scores are discussed.
Keywords: cut-off scores, standard-setting, multinomial logistic regression, polytomous logistic regression, proportional odds model.

En este artículo, las autoras derivan una ecuación general para calcular múltiples puntos de corte en la puntuación total de un test con el fin de clasificar a los individuos en más de dos categorías ordinales. La ecuación se deriva a partir del modelo de regresión logística multinomial (RLM), que es una extensión del modelo de regresión logística binaria (BLR) para variables de respuesta politómica. Con este procedimiento analítico, los puntos de corte se establecen en la puntuación del test (la variable predictora) en la que un individuo tiene la misma probabilidad de pertenecer a la categoría j que a la categoría j+1 de una variable de respuesta ordinal. La aplicación del procedimiento completo se ilustra a través de un ejemplo con datos de un estudio real sobre trastornos de la conducta alimentaria. En este ejemplo se obtienen dos puntos de corte en las puntuaciones del Test de Actitudes Alimentarias (EAT-26) para clasificar a los individuos en tres categorías ordinales: asintomático, sintomático o con trastorno de la conducta alimentaria. Los diagnósticos se obtuvieron a partir de las respuestas a un autoinforme (Q-EDD) en el que se operativizan los criterios del DSM-IV para los trastornos de la conducta alimentaria. Se discuten diferentes alternativas al modelo RLM para establecer múltiples puntos de corte.
Palabras clave: puntos de corte, establecimiento de estándares, regresión logística multinomial, regresión logística politómica, modelo de odds proporcionales.

This research was supported by grants from the Ministerio de Ciencia y Tecnología (Project Ref. BSO2001-1945) and Consejería de Educación y Ciencia de la Junta de Andalucía (Research Group CTS-278).

Correspondence concerning this article should be adressed to Rosa Bersabé Morán. Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Málaga. 29071 Málaga. (Spain). E-mail: bersabe@uma.es

In educational and psychological testing, the item responses to a test are usually transformed into a total score that is distributed within a range of values (e.g., from 0 to 50). In some decision making situations, one cut-off score must be established on the total test score in order to classify individuals into two categories (e.g., whether a psychological disorder is present or absent). In clinical settings, this type of problem is frequently solved with the construction of Receiver Operating Characteristic (ROC) curves. ROC analysis was derived from the theory of signal detectability (TSD) which was developed in the early 1950s by engineers working on problems of radar and sonar detection (Peterson, Birdsall & Fox, 1954; Van Meter & Middleton, 1954). ROC curves are used to summarize diagnostic performance of a test by plotting true-positive rate (sensitivity) versus false-positive rate (1- specificity) for each possible cut-off score (threshold). From ROC analysis, a cut-off score on a test can be established taking into account factors such as disorder prevalence (or base rate), and the cost-benefit relation of the various decisions (He, Metz, Tsui, Links & Frey, 2006).

Traditionally, ROC analysis is only used to evaluate the performance of binary classification. However, some classification tasks involve distinguishing between three or more categories and, therefore, multiple cut-off scores on a test must be established to assign individuals to one of the outcome types. Several authors have proposed generalized ROC models to be applied with more than two categories or classes (e.g., He *et al*., 2006; Mossman, 1999; Provost & Fawcett, 2001; Scurfield 1996, 1998). These multiple-class ROC models extend the basic concepts and mathematics of the traditional two-class ROC analysis. For a dichotomous diagnostic task, the true identification rates (sensitivity and specificity) can be plotted as a ROC *curve*. In multiple-class diagnostic problems, a ROC *surface* describing the true identification rates can be plotted in a multi-dimensional space. Consequently, multiple-class ROC analysis involves substantial increase in complexity compared to the two-class problem. These models are in course of development and computational calculus, applications and other questions related to these methods remain to be solved. As noted by He *et al*. (2006): "Much work remains to be done for multiple-class ROC analysis to achieve the level of maturity and utility of two-class ROC analysis" (p. 580).

In educational settings, a total test score is also frequently used to classify individuals into more than two groups (e.g., basic, proficient, advanced). Within this context, several methods to establish multiple cut-offs on the total test score have been proposed (for a review, see Cizek & Bunch, 2007). The computed cut-off scores depend on the particular standard setting method used. In the *holistic methods*, the calculation of cut-off scores is based on the relationship between the examinee papers (test scores) and the ratings they receive by one or more judges who render a single (holistic) verdict for each work sample. For example, in the *Analytic Judgment method* (Plake & Hambleton, 2001) individuals classified into a particular category are subdivided into borderline groups, and mean test scores are determined for each of these borderline groups. Cut-off scores are then computed as the midpoint between two adjacent borderline group means. For example, the cut-off score distinguishing "basic" from "proficient", was obtained by averaging the test scores of individuals classified into the "high-basic" and "low-proficient" borderline categories. Similarly, one might simply calculate the mean (or median) of the test scores for each category group, and then calculate the midpoint between two adjacent category means to derive a cut-off score (Cizek, Bunch & Koons, 2004).

In the *Body of Work method* (see Kingston, Kahl, Sweeney & Bay, 2001, pp. 230-231), binary logistic regression is used to calculate multiple cut-off scores based on the final round of ratings. Binary logistic regression (BLR) is a statistical method used to predict the probability of success based on an item of information (e.g., a test score). To establish a cut-off score, the primary goal is to identify the test score at which the likelihood of success is equal to the likelihood of failure ($p = .50$). Once the parameters of the BLR model have been estimated, the cut-off score can be directly computed as $-\alpha/\beta$, where α is the intercept, and β is the regression coefficient.

In the context of the *Body of Work method*, separate BLR models are applied by defining success as being assigned to category "*j or higher"*. For example, with three categories (e.g., basic, proficient, advanced), two separate logistic regression models would be fitted to the data to predict success. In one case, success would be defined as being assigned to category "proficient or advanced", and in another it would be defined as being assigned to category "advanced". Thus, in this analytical procedure, multiple cut-off scores are calculated by fitting separate BLR models corresponding to the sequential partitioning of the data. However, it is possible to calculate all cut-off scores simultaneously through an ordinal logistic regression model, e.g., the proportional odds (PO) model (Cizek & Bunch, 2007). The goal of the PO model is to estimate the odds of being "*at or above"* a given category across all consecutive cumulative splits to the data (O'Connell, 2006).

By using either a PO model or a BLR model in the *Body of Work method*, cut-off scores are established at the test score at which the probability of being "at or above" category *j* is the same as the probability of being "below" category *j*. It is useful to think of these cut-off scores as marking the consecutive points at which an individual might be predicted into a category *equal or greater* than *j*. Nevertheless, multiple cut-off scores can be better defined as the test scores at which an individual is as likely to be in category j as in category $j+1$. Following this definition,

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

the aim of this study is to derive an equation that allows us to obtain multiple cut-off scores at intersection of adjacent category distributions. This equation will be derived from the multinomial logistic regression model.

Multinomial logistic regression

The multinomial logistic regression (MLR) model, is also referred to as the polytomous logistic regression model. Details of the statistical theory underlying the MLR model can be found in several sources, for example, Agresti (2002), Hosmer and Lemeshow (2000), and Kleinbaum and Klein (2002). A description of the MLR model with an application using the SAS system can be found in Allison (1999), and Stokes, Davis and Koch (2000).

The MLR model is a straightforward extension of the binary logistic regression model for dichotomous outcomes, to accommodate polytomous outcome variables (*i.e.*, outcome variables with more than two ordinal or nominal categories). In the MLR model, one of the categories of the outcome variable is designated as the reference category, and each of the other categories is compared with this baseline. The choice of reference category can be arbitrary, often the last or the first one, and is at the discretion of the researcher. Changing the reference category does not change the form of the model, but it does change the values and interpretation of the parameter estimates in the model (Kleinbaum & Klein, 2002).

Let Y be a polytomous outcome variable with J categories, and let X (the test score) be the predictor variable. With only one predictor variable, the univariate MLR model represents the conditional probabilities of each outcome category ($Y = y_j$) given the value of the predictor variable (x) (Ananth & Kleinbaum, 1997, p. 1327):

$$P(Y = y_j \mid x) = \frac{\exp(\alpha_j + \beta_j x)}{\sum_{h=1}^{J} \exp(\alpha_h + \beta_h x)}$$

$$j = 1,2,\cdots,J \qquad (1)$$

where α_j are the unknown intercept parameters, and β_j are the unknown regression coefficients corresponding to X. The parameters corresponding to the reference category are equal to zero. For example, if the first category ($Y = y_1$) has been designated as the reference category, then $\alpha_1 = 0$ and $\beta_1 = 0$.

The unknown parameters are estimated by unconditional maximum likelihood (ML). The joint probability for the likelihood function is the product of all individual subject probabilities, assuming outcomes are independent. The resulting parameter estimates are consistent, asymptotically normal, and asymptotically efficient.

Once the parameters of the MLR model have been estimated, we can predict the probability of each outcome category, $\hat{P}(Y = y_j \mid x)$, which must sum to 1, for a given value of the predictor variable. It is then possible to estimate the outcome category ($\hat{Y}$) by simply assigning each individual (with a given x value) to the category with the highest probability.

The MLR model is frequently expressed in logit form. Assuming the first category as the reference category, the univariate model has the following representation:

$$\ln\left[\frac{P(Y = y_j \mid x)}{P(Y = y_1 \mid x)}\right] = \alpha_j + \beta_j x$$

$$j = 2,\cdots,J \qquad (2)$$

Exponentiating the regression coefficient, $\exp(\beta_j)$, will result in the odds ratio comparing ($Y = y_j$) with ($Y = y_1$) for a unit increase in x. Notice that the regression coefficient (β_j) is allowed to vary with j. This property implies that, unlike the proportional odds model, the MLR model does not assume the restricted condition of proportionality in the log-odds ratio.

Equation to set multiple cut-off scores

A cut-off score can be defined as the test score at which an individual is as likely to be in category j as in category $j+1$. Therefore, a cut-off score may be defined formally as the x_j value satisfying

$$P(Y = y_j \mid x_j) = P(Y = y_{j+1} \mid x_j)$$

$$j = 1,2,\cdots,J-1 \qquad (3)$$

Note that with more than two outcome categories these probabilities are not necessarily equal to .50. Replacing the previous expression with Equation (1), we get

$$\frac{\exp(\alpha_j + \beta_j x_j)}{\sum_{h=1}^{J} \exp(\alpha_h + \beta_h x)} = \frac{\exp(\alpha_{j+1} + \beta_{j+1} x_j)}{\sum_{h=1}^{J} \exp(\alpha_h + \beta_h x)}$$

(4)

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

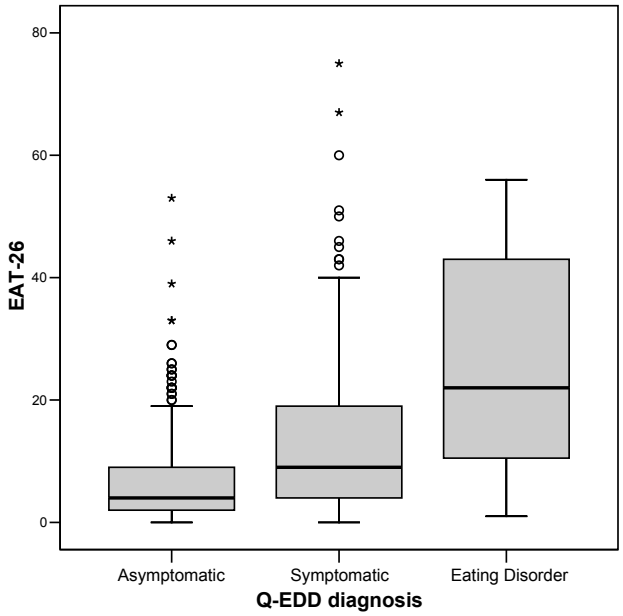

Figure 1. Box plot of EAT-26 total scores in the different groups.

Multiplying both terms of the above equation by $\sum_{h=1}^{J}\exp(\alpha_h+\beta_h x)$, yields

$$\exp(\alpha_j+\beta_j x_j)=\exp(\alpha_{j+1}+\beta_{j+1}x_j) \quad (5)$$

And, solving for x_j, the value of the cut-off score can be obtained as

$$x_j=\frac{\alpha_j-\alpha_{j+1}}{\beta_{j+1}-\beta_j} \qquad j=1,\cdots,J-1 \qquad (6)$$

where α_j and β_j are the parameters of the MLR model. Note that for a dichotomous outcome Y (usually coded as 0 and 1), the MLR and binary logistic regression models reduce to the same model. Thus, the only cut-off score of the binary logistic regression model can be calculated as: $x_1=(\alpha_1-\alpha_2)/(\beta_2-\beta_1)$, where $\alpha_1=0$ and $\beta_1=0$ (if the first category of the outcome, $y_1=0$, is designated as the reference category), and, α_2 and β_2 are the parameters corresponding to the second category of the outcome ($y_2=1$).

Example

An example to illustrate the complete application of the procedure to set two cut-off scores on a test is given. Basically, the applied procedure involves the following steps: (1) draw a representative sample of individuals from the population of test takers; (2) obtain a continuous test score (X), and an ordinal criterium score (Y) for each individual; (3) fit MLR model-data, performing goodness-of-fit tests; (4) substitute the parameter estimates of the MLR model into Equation (6) to obtain the cut-off scores; (5) assess the correspondence between observed and predicted categories, included in the classification table, by obtaining different indices of predictive efficiency.

In this example, the aim of the study is to set two cut-off scores on the EAT-26 (Garner, Olmsted, Bohr & Garfinkel, 1982), which is the short version of the *Eating Attitudes Test* (EAT-40; Garner & Garfinkel, 1979) adapted into Spanish by Castro, Toro, Salamero & Guimera (1991). Psychometric properties of this test have also been studied in different Spanish female samples (Rivas, Bersabé, Jiménez & Berrocal, in press). The EAT-26 total score (X) ranges from 0 to 78. On the other hand, the outcome (Y) has three ordered categories: (1) individuals without symptoms of an eating disorder (ED) (asymptomatics); (2) individuals presenting some symptoms but not a full ED (symptomatics); and, (3) individuals with an eating disorder (ED). These three groups were formed by the responses to the Spanish version of the *Questionnaire for Eating Disorder Diagnoses* (Q-EDD; Mintz, O´Halloran, Mulholland & Schneider, 1997) developed by Rivas, Bersabé & Castro (2001). The Q-EDD is a self-report that operationalises DSM-IV criteria for eating disorders (American Psychiatric Association, 1994). The sample of participants comprised 778 females aged 12 to 21 from different high schools in Malaga (Spain). Although the results may be informative in the field of eating disorders, the main purpose of the example is to explain the proposed procedure.

Statistical software

The MLR model was fitted using the *multinom* function in *nnet* package (Venables & Ripley, 2002) included in R environment (R Development Core Team, 2008). Multinom function fits multinomial log-linear models via neural networks. R is a language and environment for statistical computing and graphics (Venables, Smith & R Development Core, 2004). R has at least three compelling advantages: it is *free* software as part of the GNU Project; it runs on multiple platforms (e.g., Windows, Unix and Macintosh); and combines many of the most useful statistical programs into one integrated environment (Revelle, 2007).

Many other statistical packages include a procedure to carry out multinomial logistic regression: e.g., NOMREG procedure included in SPSS (2006) or PROC CATMOD included in SAS (2000), to cite but a few of the widely used software packages.

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

Table 1
Parameter estimates: Results of fit of multinomial logistic regression model

Outcome Category [a]	Parameter	Estimate	SE	Wald statistic
Symptomatic	α_2	-1.592	.123	-12.954
	β_2	.080	.010	8.037
Eating Disorder	α_3	-4.257	.303	-14.029
	β_3	.127	.014	9.282

a Reference category : Asymptomatic ($\alpha_1 = 0$ and $\beta_1 = 0$)

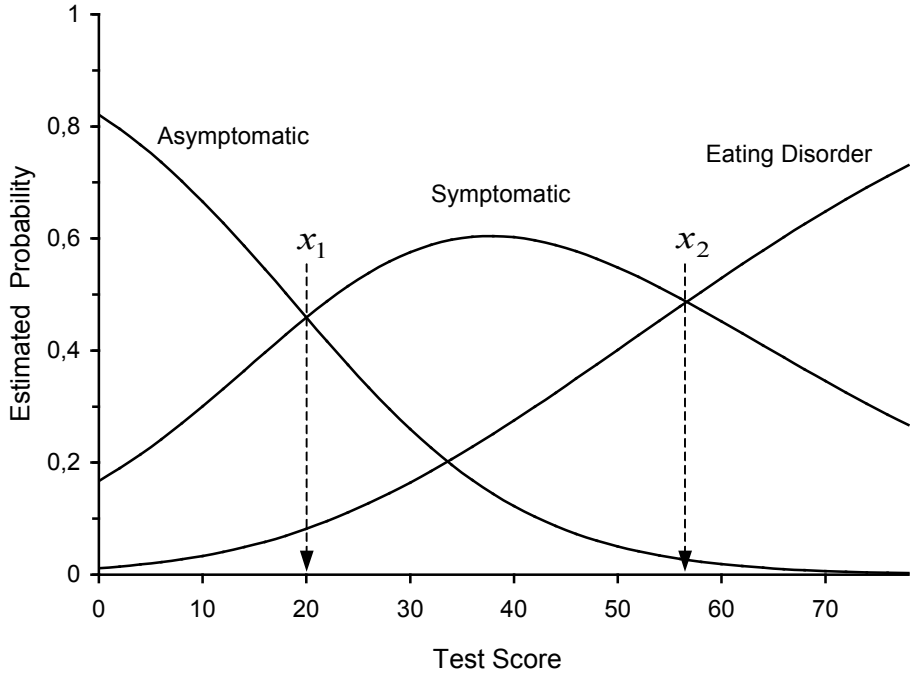

Figure 2. Predicted conditional probabilities of being in each category given the test score.

Results

Prior to fitting the MLR model, it is convenient to make a descriptive analysis in order to show the relationship between the test scores and the outcome categories. For this purpose, females were classified into one of the following groups: Asymptomatic ($N = 525$), Symptomatic ($N = 217$), or Full Eating Disorder ($N = 36$), according to the responses on the Q-EDD. In each group, the distribution of the total test scores (EAT-26) is represented by a box plot (Figure 1). The EAT-26 scores show an increasing trend over the three diagnostic groups, from the lowest scores (asymptomatic group) to the highest (eating disorder group). Therefore, it would seem that the three diagnostic categories can be ordered as a function of the total test scores.

In this study, the outcome variable (Y) is the diagnosis obtained from responses to the Q-EDD, with three ordinal levels (coded 1 for asymptomatic, 2 for symptomatic, and 3 for eating disorder). The only predictor variable (X) is the test (EAT-26) score. These two variables were analyzed running a MLR model. The goodness-of-fit of the model was assessed by comparing its residual deviance ($D_m = -2LL_m = 1056.702$) with the residual deviance of

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

Table 2
Classification table

Observed category (Q-EDD diagnosis)	Predicted category $0 \le x \le 20$ Asymptomatic	$21 \le x \le 56$ Symptomatic	$57 \le x \le 78$ Eating disorder	Total
Asymptomatic	503	22	0	525
Symptomatic	171	43	3	217
Eating Disorder	18	18	0	36
Total	692	83	3	778

the null model ($D_0 = -2LL_0 = 1188.409$), which includes only the intercepts. The deviance is a measure of how *poorly* the model reproduces the observed data. The likelihood ratio test ($G = D_0 - D_1 = 131.707$, $df = 2$, $p < .001$) compares these two deviances. The null hypothesis is rejected, indicating a statistically significant decrease in the deviance when the predictor (X) is included in the model. This means that the model fits the data better than the null model, in terms of correspondence between observed and predicted conditional probabilities. The reduction in deviance can be expressed through the likelihood ratio $R_L^2 = 1 - (D_m / D_0) = .111$. For this model, the inclusion of the predictor in the model reduces the deviance of the null model by approximately 11%.

The parameter estimates of the MLR model are given in Table 1. For y_1 (the reference category), the model given in Equation (1) represents the conditional probability of being in the asymptomatic category, given the test score (x):

$$P(Y = y_1 \mid x) = \frac{1}{1 + \exp(\alpha_2 + \beta_2 x) + \exp(\alpha_3 + \beta_3 x)} \quad (7)$$

For y_2, the model gives the conditional probability of being in the symptomatic category:

$$P(Y = y_2 \mid x) = \frac{\exp(\alpha_2 + \beta_2 x)}{1 + \exp(\alpha_2 + \beta_2 x) + \exp(\alpha_3 + \beta_3 x)} \quad (8)$$

And, for y_3, the model gives the conditional probability of being in the eating disorder category:

$$P(Y = y_3 \mid x) = \frac{\exp(\alpha_3 + \beta_3 x)}{1 + \exp(\alpha_2 + \beta_2 x) + \exp(\alpha_3 + \beta_3 x)} \quad (9)$$

Figure 2 shows three curves with the estimated conditional probabilities for each diagnostic category, $\hat{P}(Y = y_j \mid x)$. As can be seen, the higher the test scores, the lower the probability of being in the asymptomatic category and, conversely, the higher the probability of having a full eating disorder.

In Figure 2, the two cut-off points are represented graphically as the test score corresponding to the intersection of the adjacent curves (1 with 2, and 2 with 3), *i.e.*, the test score (x_j) at which an individual is as likely to be in category j as in category $j+1$, for $j = 1,2$. These cut-off scores can be obtained by applying Equation (6) to these data:

$$x_1 = -\alpha_2 / \beta_2 = 20.003$$

$$x_2 = \frac{\alpha_2 - \alpha_3}{\beta_3 - \beta_2} = 56.635$$

Given the obtained value for the first cut-off score, and applying Equation 1, it can be shown that the estimated conditional probability of being in the "asymptomatic" category is equal to the estimated probability of being in the "symptomatic" category: $\hat{P}(Y = y_1 \mid x_1) = \hat{P}(Y = y_2 \mid x_1) = 0.459$. Similarly, it can be shown that the estimated probability of being in the "symptomatic" category is equal to the probability of being in the "eating disorder" category given the obtained value for the second cut-off score: $\hat{P}(Y = y_2 \mid x_2) = \hat{P}(Y = y_3 \mid x_2) = 0.487$. This proves that the obtained cut-off scores have been clearly established at intersection of adjacent distributions.

Based on the obtained cut-off scores, the asymptomatic category ($\hat{Y} = 1$) is estimated for the test scores (x) within the 0-20 interval; the symptomatic category ($\hat{Y} = 2$), for the x values from 21 to 56; and the eating-disordered category ($\hat{Y} = 3$), for x values from 57 to 78. Table 2 shows the results of the classification, comparing the Q-EDD diagnoses (observed outcome category, Y) with those estimated as a function of the EAT-26 scores (predicted outcome category, $\hat{Y}$). In this example, the overall accuracy rate was 70.18%. In the asymptomatic group, the EAT-26 correctly classified 95.8 % of individuals, and 4.2% were erroneously classified as symptomatic. In the symptomatic group, 19.8% of individuals were correctly classified from the test score; 78.8% were erroneously

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

classified as asymptomatic; and 1.4% as eating disorder. In the eating disordered group, none of the individuals was correctly diagnosed, 50% were erroneously classified as asymptomatic, and 50% as symptomatic.

Other indices of predictive efficiency assess the correspondence between observed and predicted categories included in the classification table. Some of them can be used with polytomous outcome variables (for a review, see Menard, 1995). These include λ_p, also called *adjusted count* R^2, which is defined (Long, 1997; Menard, 2000) as $\lambda_p = 1-(n-\sum f_{ii})/(n-n_{\text{mode}})$, where n is the sample size, f_{ii} is the number of cases for which the predicted value is equal to the observed value, and n_{mode} is the observed number of cases in the modal category of the dependent variable (maximum row marginal). In this example (Table 2), $\lambda_p = 1-232/253 = .083$, which means that predicting outcome category using the MLR model, decreases the initial prediction error by 8.3%. However, this reduction of the initial prediction error is not statistically significant, as indicated by the normal approximation to the binomial test ($d = 1.607 \quad p = .054$). Initial prediction error is defined as the expected errors of the null model (without the predictor). For this index, the null prediction model uses the mode of the dependent variable as the predicted value for all cases.

Another index of predictive efficiency is τ_p which is defined (Menard, 2000) as $\tau_p = 1-(n-\sum f_{ii})/[\sum f_i(n-f_i)/n]$, where f_i is the observed frequency for category i. For the actual classification table, $\tau_p = 1-232/361.535 = .358$ showing that initial classification error is reduced by approximately 36% using the MLR model. This reduction of the initial classification error is statistically significant, as indicated by the normal approximation to the binomial test ($d = 1.607$, $p = .054$). Initial classification error is defined as the expected errors of the null model (without the predictor). For this index, the null model requires that cases must be classified into distinct categories adjusted to the observed marginal distribution of the dependent variable (base rates).

Hosmer and Lemeshow (2000) point out that model fit in terms of correspondence between observed and estimated probabilities (e.g. likelihood ratio test) is often more reliable and meaningful than assessment of fit based on classification. They suggest that classification statistics be used as an adjunct to other measures, rather than a sole indicator of quality of the model. Indeed, a well fitted model in terms of residual deviance may not necessarily lead to high predictive efficiency rates (*i.e.*, classification rates), as shown in the present example.

Discussion

A general equation to compute multiple cut-off scores on a test has been derived from the MLR model. From this analytical procedure, a cut-off score has been defined as the test score at which an individual is as likely to be in category j as in category $j+1$. In other words, a cut-off score can be defined as the abscissa (test score) corresponding to the intersection of adjacent category distributions.

Following this definition, multiple cut-off scores can also be obtained by fitting separate binary logistic regression (BLR) models for each pair of adjacent categories. For example, with three categories (e.g., asymptomatic, symptomatic, eating disorder), two separate BLR models would be fitted to the data to predict success. In one case, success would be defined as being assigned to "symptomatic" category (versus "asymptomatic"), and in another it would be defined as being assigned to "eating disorder" category (versus "symptomatic").

Using separate BLR models yields the same *number* of parameters as those obtained by using a MLR model (e.g., two intercepts and two slopes for a three-category outcome). However, the likelihood function for the MLR model utilizes the data involving all categories of the outcome variable in a single function. In contrast, the likelihood function for a BLR model uses the data involving only two categories of the outcome variable. That is, different likelihood functions are used when fitting each dichotomous model separately than when fitting a MLR model that considers all categories simultaneously. Consequently, both the estimation of the parameters and the estimation of the variances of the parameter estimates may differ when comparing the results from fitting separate dichotomous models to the results from the MLR model (Kleinbaum & Klein, 2002).

Another option to obtain multiple cut-off scores at intersection of adjacent distributions is to use an ordinal logistic regression model, e.g., the proportional odds (PO) model (Bersabé, Rivas & Berrocal, 2009). There are certain advantages to be obtained through the fitting of a PO model, as well as certain caveats to keep in mind. The PO model summarizes the relationship between the ordinal outcome and the independent variable with only one beta parameter (β), implying that the model assumes that the effect of X on Y is the same regardless of the category of the outcome variable (j). This assumption of the model is called the proportional odds assumption, and hence the name *proportional odds* model (McCullagh, 1980). The PO model provides a parsimonious regression model for ordinal response variables. However, only if the assumption of proportional odds is tenable should the PO model be applied (Bender & Grouven, 1998).

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

In contrast, the MLR model does not assume the proportional odds condition. In the MLR model (unlike the PO model), the regression coefficient (β_j) depends on *j*, implying that the model does not assume that the effect of X on Y is the same regardless of *j* (Ananth & Kleinbaum, 1997). In this sense, the MLR model is less restrictive than the PO model, and it can be applied whether the proportional odds assumption is satisfied or not.

In an educational context, other methods involving a lower level of statistical complexity have been proposed to calculate multiple cut-off scores. For example, in the *Analytic Judgment method* (Plake & Hambleton, 2001) individuals classified into a particular category are subdivided into borderline groups, and cut-off scores are simply computed as the midpoint between two adjacent borderline group means. Because extreme scores may unduly influence a mean, the median may be used instead. This approach has the advantage of being fairly straightforward. It is easy to explain to the public and also to calculate (*i.e.*, only simple means -or medians- of test scores assigned to borderline categories are needed). However, this method has the weakness that it does not use all the available data, sacrificing information present in the test scores for non-boundary categories.

Another option is to simply calculate the mean (or median) of the test scores for each complete category group (*i.e.*, individuals classified into a particular category are not subdivided into borderline groups), and then to calculate the midpoint between two adjacent category means to derive a cut-off score (Cizek, Bunch & Koons, 2004). This option has appeal in that it uses all the available information. Similarly, the MLR model does not need to sacrifice any data from the sample of participants. However, the sophistication of the MLR analysis requires access to statistical software and a level of statistical training that may challenge some practitioners. It is definitely more difficult to explain MLR model fitting to the public than would be simple averages.

In any case, multinomial logistic regression provides a sound statistical model whose goodness-of-fit can be assessed in terms of residual deviance, which allows for the statistical comparison of different models. In contrast, methods based on simple measures of central tendency can only assess cut-off point efficiency in terms of the correspondence between observed and predicted categories included in the classification table, which is often less reliable and meaningful (Hosmer & Lemeshow, 2000). Furthermore, MLR model should yield more accurate cut-off points than methods based on a midpoint between two means (or medians), especially when the rates by category groups are very different.

A limitation of the MLR model to establish multiple cut-off scores is that, unlike ROC models, it does not take into account the proportion of the outcome categories at the population level (base rate). For this reason, the proposed methodology should only be used with randomly sampled data that resemble the base rate from the corresponding population.

Furthermore, in the suggested procedure, the costs of different classification errors are assumed to be equal. In contrast, in some of the proposed three-class ROC models (e.g., He *et al*., 2006), the individuals can be classified by making the decision that provided the maximal expected utility of incorrect decisions relative to the other two. Nevertheless, when there are more than two outcome categories, it is extremely difficult to provide realistic assessments of the relative costs of the different kinds of misclassification.

In conclusion, the MLR model provides a *joint* model that can be used to compute multiple cut-offs on a continuous test score, at intersection of adjacent distributions. The proposed methodology is recommended whenever the data represent a random sample from the corresponding population, and the relative costs of incorrect decisions are assumed to be equal. Additional research is needed to compare results using alternative procedures. It would contribute to drawing a firmer conclusion regarding comparability of the different analytical approaches.

References

Agresti, A. (2002). *Categorical Data Analysis,* 2nd ed. New York: John Wiley & Sons.

Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application*. Cary, NC: SAS Institute Inc.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology, 26*, 1323-1333.

Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology, 51,* 809-816.

Bersabé, R., Rivas, T., & Berrocal, C. (2009). Obtaining equations from the proportional odds model to set multiple cut scores on a test. *Methodology, 5*, 123-130.

Castro, J., Toro, J., Salamero, M., & Guimera, E. (1991). The Eating Attitudes Test: Validation of the Spanish version. *Psychological Evaluation, 7*(2), 175-189.

Cizek, G. J., & Bunch, M., B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M., B. & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*, 31-50.

https://doi.org/10.1017/S1138741600004042 Published online by Cambridge University Press

Garner, D.M., & Garfinkel, P.E. (1979). The Eating Attitudes Test: An index of the symptoms of anorexia nervosa. *Psychological Medicine, 9*, 273-279.

Garner, D.M., Olmsted, M.P., Bohr, Y., & Garfinkel, P.E. (1982). The Eating Attitudes Test: psychometric features and clinical correlates. *Psychological Medicine, 12*, 871-878.

He, X., Metz, C. E., Tsui, B. M. W., Links, J. M., & Frey, E. C. (2006). Three-Class ROC Analysis-A Decision Theoretic Approach Under the Ideal Observer Framework. *IEEE Transactions on Medical Imaging, 25*, 571-581.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Kleinbaum, D. G., & Klein, M. (2002). *Logistic Regresión. A Self-Learning Text,* 2nd ed. London: Springer.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109-142.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17-24.

Mintz, L. B., O'Halloran, M. S., Mulholland, A. M., & Schneider, P. A. (1997). Questionnaire for Eating Disorder Diagnoses: Reliability and validity of operationalizing DSM-IV criteria into a self-report format. *Journal of Counseling Psychology, 44*, 63-79.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making, 19*, 78-89.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks, CA: Sage.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group of Information Theory, 4*, 171-212.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning, 42*, 203-231.

R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Retrieved April 18, 2008, from: http://www.R-project.org/.

Revelle, W. (2007). *Using R for psychological research: A simple guide to an elegant package*. Retrieved June 21, 2008, from: http://www.personality-project.org/r/r.guide

Rivas, T., Bersabé, R., & Castro, S. (2001). Propiedades psicométricas del Cuestionario para el Diagnóstico de los Trastornos de la Conducta Alimentaria (Q-EDD). *Psicología Conductual, 9,* 255-266.

Rivas, T., Bersabé, R., Jiménez, M., & Berrocal, C. (in press). The Eating Attitudes Test (EAT-26): Reliability and Validity in Spanish Female Samples. *The Spanish Journal of Psychology.*

SAS Institute, Inc (2000). *SAS/STAT User's Guide, Version 8.0*. Cary, NC: Author.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology, 40*, 253-269.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to *n*-event *m*-dimensional forced-choice tasks. *Journal of Mathematical Psychology, 42*, 5-31.

SPSS, Inc (2006). *SPSS 14.0 Base User's Guide.* Chicago, IL: Author.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000).*Categorical Data Analysis Using the SAS System,* Second Edition. Cary, NC: SAS Institute Inc.

Van Meter, D., & Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group of Information Theory, 4*, 119-145.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S,* 4th ed. New York: Springer. Retrieved May 5, 2008, from: http://www.stats.ox.ac.uk/pub/MASS4.

Venables, W. N., Smith, D. M., & R Development Core (2004). *An Introduction to R*. Network Theory Limited. Retrieved June 7, 2008, from: http://cran.r-project.org/manuals.html

Received December 15, 2008
Revision received July 14, 2009
Accepted July 30, 2009