PROPOSAL

1. MULTIPLE CHOICE TESTS

1.1. Introduction

We are considering multiple choice tests (MCT) with n items with K alternatives each. The number of alternative answers is always the same for all the items of the test and there is only one correct alternative, therefore, the remainder K–1 options are distractors. The raw data must be summarized as a contingency as shown in Table 1, where x_{ij} is the number of times that the answer given is alternative j when the correct one is alternative i; r_i is the number of times that the correct alternative is i; c_j is the number of times that the student chooses the alternative answer j.

Table 1

		Student's answer			
		1	⋯	K	Totals
Right answer	1	x_{11}	⋯	x_{1K}	r_1
	⋮	⋮	⋱	⋮	⋮
	K	x_{K1}	⋯	x_{KK}	r_K
	Totals	c_1	⋯	c_K	n

1.2. Sampling

The model assumes that the number of times r_i that the correct alternative is i, is previously fixed by the teacher (gold standard in rows). Hence, the sampling is always of type II (K independent multinomial distributions).

1.3. Response model

We assume that with an MCT the probability of choosing option j when i is the correct one is given by

$$p_{ij} = \delta_{ij}\Delta + (1-\Delta)\pi_j \qquad (i, j = 1,\ldots,K)$$

where δ_{ij} is the Kronecker delta, π_j are the probabilities of the subject choosing the alternative in position j when the answer is not known and is guessed, and Δ is the parameter of interest (the proportion of items that, being recognised by the examinee, are correctly answered not due to chance).

Hence, p_{ii} represents the probability of adequately choosing those questions where the correct option is the one occupying position i, while in the opposite case, that

is when $i \neq j$, p_{ij} represents the probability of choosing a distractor.

The measure of knowledge Δ was proposed firstly by Lord and Novick (1968), and later it was considered by Hutchinson (1982) and Martín and Luna (1989, 1990). Martín and Luna considered the conditional estimation of Δ. The actual program, gives both the conditional and the unconditional –based on the maximum likelihood principle– estimations of Δ.

For the unconditional model, one-sided and two-sided confidence intervals are given. These CI are obtained by both, the classic Wald method and the test inverted method, which was proposed by Agresti and Min (2001).

This model is part of a family of models previously developed by the authors Martín and Femia (2004, 2005, 2008). The full development of the actual model can be found in Femia and Martín (2011).

2. NOMINAL AGREEMENT MODEL

2.1. Introduction

Let two raters ($R \equiv$ rows and $C \equiv$ columns) independently classify n subjects within K nominal categories. Given a subject, rater R classifies it as belonging to type i (event R_i: i = 1, 2, ..., K) and rater C as belonging to type j (event C_j: j = 1, 2, ..., K), which gives a table of frequencies as shown in Table 2. The aim is to obtain measures of agreement or concordance among R and C.

Table 2
Frequencies observed when two raters (R and C) classify n subjects in K categories

R/C Answer	C_1	. . .	C_j	. . .	C_K	Total
R_1	x_{11}	. . .	x_{1j}	. . .	x_{1K}	r_1
.	.		.		.	.
R_i	x_{i1}	. . .	x_{ij}	. . .	x_{iK}	r_i
.	.		.		.	.
R_K	x_{K1}	. . .	x_{Kj}	. . .	x_{KK}	r_K
Total	c_1	. . .	c_j	. . .	c_K	n

2.2. Response model

For this purpose, there are two aspects of interest that define the adopted model:

Samples: The data can have been obtained in two different ways: Sampling I (r_i and c_j have been obtained at random ≡ a multinomial distribution of size n) or

Sampling II (r_i are previously fixed ≡ K independent multinomial distributions of sizes r_i).

Raters: There are three possibilities: *R* is a gold standard rater; *C* is a gold standard rater; neither *R* nor *C* are gold standard raters.

Martín and Femia (2004) propose the following answer model:

$$\Pr(C_i|R_i) = \Delta_i+(1-\Delta_i)\pi_i = p_{ii}\ ,\ \Pr(C_j|R_i) = (1-\Delta_i)\pi_j = p_{ij}\ (j\neq i),\quad P(R_i \cap C_j)=q_i p_{ij}$$

where:

$$0 \le \pi_j \le +1\ (\Sigma\pi_j =1),\quad \Delta_i \le +1,\quad 0 \le p_{ii} \le +1,\quad 0 \le p_{ij} \le +1,\quad 0 \le q_i \le +1$$

The interpretation of the parameters in the model is:

π_j = Proportion of objects R_i that, not being recognized by rater *C*, are classified as C_j due to chance.

Δ_i = Proportion of objects R_i that, being reconognized by rater *C*, are classified as C_i not due to chance. This value may be negative when *C* recognizes the object R_i "the wrong way round".

q_i = Proportion of objects of type R_i (in sampling I).

2.3. Measures of agreement

The measures of agreement crude in the left-hand part of Table 3 are the traditional and intuitive ones, although they are not valid in every case but depend on the Model (see Table 4). However these measures have the disadvantage of not taking the effect of chance into account. In the case of the Overall Agreement it is traditional to correct this effect by using Cohen's *coefficient of concordance* κ (kappa) (1960). The program gives the values for $\hat{\kappa}$ and for S.E.($\hat{\kappa}$).

As a result, the part of $x_{ii}=r_i\,\hat{p}_{ii}$ that is not due to chance is $r_i\hat{\Delta}_i$, which yields the estimators for the right-hand part of Table 2. Martín and Femia (2004, 2005 and 2008) define these parameters and obtain their S.E. Note that the parameter $\hat{\mathcal{F}} =\hat{\mathcal{P}} =\hat{\mathcal{A}} =\hat{\mathcal{S}} =\hat{\Delta}$ is an alternative to $\hat{\kappa}$ because it is valid as a measure of *conformity, consistency and agreement (concordance)*. Moreover, Martín and Femia (2004, 2005 and 2008) shows that Delta does not have the drawbacks of Kappa. However the new indices are not reliable in a 2×2 table when the overall value $\hat{\Delta}$ is very similar to $\hat{\Delta}_{ind} = \left\{\sqrt{x_{11}} - \sqrt{x_{22}}\right\}^2 / n$ and the marginal are very unbalanced in the same direction. In such a

case one has to assume that the data are independent (and therefore the overall agreement is null).

The program gives the estimators of maximum likelihood for π_j and Δ_i, and the estimators (and values of S.E.) of all the new parameters of Table 3.

Table 3

Measures of agreement in Category *i* (first 4 rows) and Overall (last row)

Not chance-corrected		**Chance-corrected**	
$\hat{F}_i = \frac{x_{ii}}{r_i}$	Sensitivity (Category i)	$\widehat{\mathcal{F}}_i = \hat{\Delta}_i$	*Conformity* (Category i)
$\hat{P}_i = \frac{x_{ii}}{c_i}$	Predictive Value (Category i)	$\widehat{\mathcal{P}}_i = \frac{r_i \hat{\Delta}_i}{c_i}$	*Predictivity* (Category i)
$\hat{A}_i = \frac{x_{ii}}{n}$	Agreement (Category i)	$\widehat{\mathcal{A}}_i = \frac{r_i \hat{\Delta}_i}{n}$	*Agreement/Concordance* (Category i)
$\hat{S}_i = \frac{2x_{ii}}{r_i + c_i}$	Index of Agreements (Category i)	$\widehat{\mathcal{S}}_i = \frac{2r_i \hat{\Delta}_i}{r_i + c_i}$	*Consistency* (Category i)
$\hat{F}=\hat{P}=\hat{A}=\hat{S}=\frac{\Sigma x_{ii}}{n}$	Agreement (Overall)	$\widehat{\mathcal{F}} = \widehat{\mathcal{P}} = \widehat{\mathcal{A}} = \widehat{\mathcal{S}} = \hat{\Delta} = \frac{\Sigma r_i \hat{\Delta}_i}{n}$	*Agreement* (Overall)

Table 4

Licit measures with respect to the adopted Model

	R = Standard	C = Standard	There is not a standar
Sampling II	$\widehat{\mathcal{A}}_i$, $\widehat{\mathcal{F}}_i$ and $\hat{\Delta}$	$\widehat{\mathcal{A}}_i$ and $\hat{\Delta}$	$\widehat{\mathcal{A}}_i$ and $\hat{\Delta}$
Sampling I	$\widehat{\mathcal{A}}_i$, $\widehat{\mathcal{F}}_i$, $\widehat{\mathcal{P}}_i$ and $\hat{\Delta}$	Put the standard in rows	$\widehat{\mathcal{A}}_i$, $\widehat{\mathcal{S}}_i$ and $\hat{\Delta}$

REFERENCES

Agresti, A. and Min, Y. (2001). On Small Sample Confidence Intervals for Parameters in Discrete Distributions. *Biometrics* 57, 963-971.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 37-46.

Femia Marzo, P. and Martín Andrés, A. (2014). Multiple Choice Tests: Inferences based on estimators of maximum likelihood. *Open Journal of Statistics 4*, 466-483. DOI: 10.4236/ojs.2014.46045.

Hutchinson, T. P. (1982). Some theories of performance in multiple-choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology 35*, 71-89.

Lord, F.M. and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Menlo Park, Ca: Addison-Wesley.

Martín Andrés, A. and Femia Marzo, P. (2004). Delta: A New Measure of Agreement Between Two Raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1-19.

Martín Andrés, A. and Femia Marzo, P. (2005) Chance-Corrected Measures of Reliability and Validity in K×K Tables. *Statistical Methods in Medical Research 14*, 473-492.

Martín Andrés, A. and Femia Marzo, P. (2008). Chance-corrected measures of reliability and validity in 2×2 tables. *Communications in Statistics-Theory and Methods 37*, 760-772.

Martín Andrés, A. and Luna del Castillo, J.D. (1989) Tests and Intervals in Multiple Choice Tests: a Modification of the Simplest Classical Model. *British Journal of Mathematical and Statistical Psychology 42*, 251-263.

Martín Andrés, A. and Luna del Castillo, J.D. (1990). Multiple Choice Tests: Power, length and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology 43,* 57-71.