frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 03 October 2013
doi: 10.3389/fpsyg.2013.00694

Multinomial tree models for assessing the status of the reference in studies of the accuracy of tools for binary classification

Juan Botella*, Huiling Huang and Manuel Suero

Department of Social Psychology and Research Methods, Facultad de Psicología, Universidad Autónoma de Madrid, Madrid, Spain

Edited by:
Michel Regenwetter, University of Illinois at Urbana-Champiagn, USA

Reviewed by:
Thorsten Meiser, University of Mannheim, Germany
Edgar Erdfelder, University of Mannheim, Germany

****Correspondence:***
Juan Botella, Facultad de Psicología, Universidad Autónoma de Madrid, Campus de Cantoblanco, c/Ivan Pavlov, 6, 28049 Madrid, Spain
e-mail: juan.botella@uam.es

Studies that evaluate the accuracy of binary classification tools are needed. Such studies provide 2 × 2 cross-classifications of test outcomes and the categories according to an unquestionable reference (or gold standard). However, sometimes a suboptimal reliability reference is employed. Several methods have been proposed to deal with studies where the observations are cross-classified with an imperfect reference. These methods require that the status of the reference, as a gold standard or as an imperfect reference, is known. In this paper a procedure for determining whether it is appropriate to maintain the assumption that the reference is a gold standard or an imperfect reference, is proposed. This procedure fits two nested multinomial tree models, and assesses and compares their absolute and incremental fit. Its implementation requires the availability of the results of several independent studies. These should be carried out using similar designs to provide frequencies of cross-classification between a test and the reference under investigation. The procedure is applied in two examples with real data.

Keywords: binary classification, gold standard, multinomial tree models, imperfect reference, diagnostic accuracy

INTRODUCTION

Tools for binary classification are regularly used in psychology, as in screening processes for the early detection of certain disorders or risk factors for such disorders. Their objective is to detect a specific status and assist in the decision-making process. Procedures that are able to identify a specific status quite accurately are available, but they are expensive and consequently large scale applications are unfeasible. Therefore, psychologists and other health professionals are looking for alternative classification tools which are simple, effective, and inexpensive. Often these alternatives are questionnaires that contain multiple items, the scores of which generate a dichotomy based on a simple rule that the practice suggests as an effective screening. For example, it has been proposed to use the Alcohol Use Disorders Identification Test (AUDIT) with a cut-off of $X \geq 8$ for detecting alcohol-use disorders (Babor et al., 2001) or the Mini-Mental State Examination (MMSE; Folstein et al., 1975) with a cut-off of $X \geq 24$ to detect dementia.

The effectiveness of those tools is estimated by studies that assess their accuracy in classification. These studies require a previous classification by a reference (R) that is considered unquestionable and which provides the true status of each participant. The results are summarized in 2 × 2 tables with the frequency of participants that are positive for the condition sought (denoted by 1) and of those who are not (denoted by 0) according to R, crossed with the result (positive or negative) of the test (T) for which the diagnostic accuracy is going to be assessed. For example, if the AUDIT test is applied, respectively, to groups of individuals showing alcohol abuse (N_1) and without alcohol abuse (N_0), the status is crossed with the binary classification by the test, and a table similar to that in **Figure 1** is generated. The table represents the scenario of evaluation, where the joint frequencies of the results of R and T are summarized. The four events are represented as TP (true positives), FN (false negatives), FP (false positives), and TN (true negatives). The sum of the four frequencies equals the total sample size of participants (M). The empirical prevalence in the study is the proportion of participants with the status "1" in the study, $(TP + FN)/M = N_1/M$.

The diagnostic accuracy of a binary classification test can be summarized by two probabilities: the probability of a positive result given the status "1," $P(T = 1|S = 1)$, and the probability of a negative result given the status "0," $P(T = 0|S = 0)$. In a perfect performance test both probabilities will equal 1 and would provide a contingency table where $FN = FP = 0$. However, in practice, the tests for which the accuracy is assessed have suboptimal reliability, and the probability $P(T = 1|S = 1)$, known as the *sensitivity* of the test, will be less than 1. Similarly, the probability $P(T = 0|S = 0)$, known as the *specificity* of the test, will also be less than 1. The sensitivity and specificity of the test T are denoted here by Se_T and Sp_T, respectively.

WHEN THE REFERENCE IS NOT A GOLD STANDARD

The traditional design of studies evaluating the accuracy of screening tests implies that R is a device of perfect accuracy; that is why it is called *gold standard* (GS). However, sometimes R also has a suboptimal reliability and is therefore not a gold standard. In these cases it is referred to as an *imperfect reference* (IR). Some authors have highlighted that difficulties arise when the reference is imperfect, the most important being that the estimates of sensitivity and specificity are biased, as is the calculation of the prevalence (e.g., Valenstein, 1990). With regard to the prevalence, if the study is performed with an IR, the difference between the *observed* (or apparent) *prevalence*, obtained from the frequencies in **Figure 1** and defined above as $(TP + FN)/M$, and the *empirical prevalence* (which is estimated to fit the models) must be explicitly highlighted. When the reference is imperfect the observed prevalence is calculated from "contaminated" frequencies, whereas the empirical prevalence is the (unknown) actual proportion of targets in the study.

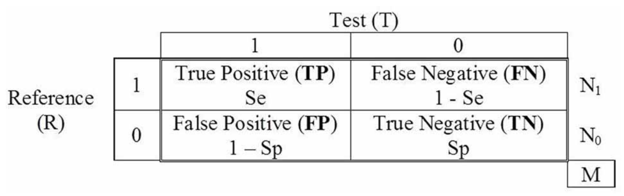
When the reference is a GS the observed prevalence and the empirical prevalence are identical, but when it is an IR they may be very different. By fitting IR models the parameter reflecting the prevalence, π, can be estimated to give an approximation to the empirical prevalence of the study (the actual proportion of participants with status "1"). The use of an IR allows some traffic of counts between FP and TP on the one hand, and TN and FN on the other. This traffic generates the difference between the empirical prevalence and the observed prevalence. Some solutions have been proposed to improve the performance of studies when R is imperfect (Rutjes et al., 2007; Reitsma et al., 2009; Trikalinos and Balion, 2012). However, before you can apply them it is necessary to assume whether R is flawed or not (whether R shows sensitivity and/or specificity lower than 100%).

Our main objective is to propose a procedure to help with the choice between two different scenarios, related to the status of a reference as a GS or an IR. The procedure consists of fitting two nested multinomial tree models (MTM) and assessing their goodness-of-fit.

MULTINOMIAL TREE MODELS

The MTM employed here is imported from the general framework of the multinomial processing tree models. This class of models is widely used in psychology for the study of cognitive processes (Batchelder, 1998; Batchelder and Riefer, 1999; Erdfelder et al., 2009). In the present application, cognitive processes (the discrete cognitive states generated by those processes, or the responses generated by those processes) are substituted by the result of administering two tools for categorizing the status of the participants in a given study.

We have used MultiTree (Moshagen, 2010), a software specifically developed for that class of models. Parameter estimation proceeds by employing the expectation maximization (EM) algorithm (Hu and Batchelder, 1994). Several statistics are available for assessing both the absolute and incremental fitting of the models (García-Pérez, 1994; Hu and Batchelder, 1994). For the goodness-of-fit test of a model, the degrees of freedom are defined as the difference between the number of independent data categories and the number of parameters estimated. For the incremental fitting, the degrees of freedom are the difference between the degrees of freedom of the two models being compared. The pattern of results of the goodness-of-fit tests for the absolute fit of both models and the incremental fitting, as nested models, will be the basis for selecting one of the models (Moshagen and Hilbig, 2011).



		Test (T)		
		1	0	
Reference (R)	1	True Positive (**TP**) Se	False Negative (**FN**) 1 - Se	N_1
	0	False Positive (**FP**) 1 – Sp	True Negative (**TN**) Sp	N_0
				M

FIGURE 1 | Contingency table of the binary classifications of a reference (R) and a test (T).

Given a single data set, several of the models considered here are not identifiable, and/or, testable. The reason is that the number of estimated parameters exceeds (or is equal to) the number of independent categories. The solution we propose to this problem can only be applied if some minimum number of independent and homogeneous studies that provide data regarding the classification of the test (and have used the same reference) are available. It is assumed that all those studies share the same parameters of accuracy, but each has its own parameter of prevalence. In the two examples with real data described below four independent studies are employed.

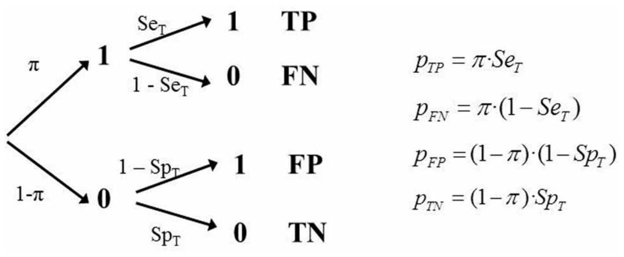
MODELS OF ASSESSMENT

Two models of assessment are involved in the present research. In the assessment scenario involved in *model 1* the reference is a GS, whereas in *model 2* it is an IR.

MODEL 1: THE REFERENCE IS A GOLD STANDARD

The study assesses the diagnostic accuracy of a test where the sampling model generates a parametric prevalence (proportion of participants with status "1" in the population, according to the study's sampling model) equal to π. The process of assessment can be represented by a tree diagram as in **Figure 2**.

Although it is possible to fit the model with only one study, testing the model requires more than one independent study that provides data regarding the classification test (and have used the same reference). A study provides three independent frequencies (the fourth is determined by M) therefore to fit the model three parameters must be estimated. And consequently if we do so there would be no degrees of freedom available to test it. However, if there are several independent studies using the same classification tool, and the same criterion, it is possible to fit and test the model properly, assuming that the sensitivity and specificity are independent of the study's empirical prevalence.



$$p_{TP} = \pi \cdot Se_T$$
$$p_{FN} = \pi \cdot (1 - Se_T)$$
$$p_{FP} = (1 - \pi) \cdot (1 - Sp_T)$$
$$p_{TN} = (1 - \pi) \cdot Sp_T$$

FIGURE 2 | Tree diagram of Model 1 (GS) and the corresponding equations.

It is assumed that results are available from a set of k independent studies, each with a different sampling model, and therefore with a different empirical prevalence, π_i; in each study there are different total sample sizes. Each study provides four frequencies with three degrees of freedom. The number of independent parameters to be estimated is $k + 2$. The number of degrees of freedom is $3k - (k + 2) = 2k - 2$. Consequently, the minimum number of studies required to test this model is 2 because then the parameters will be two π values plus Se_{T} and Sp_{T}. The degrees of freedom would be 2. It is defined a separate tree for each study and a simultaneous fit is performed. Naturally, the estimate would be more reliable the larger the number of independent estimates (k).

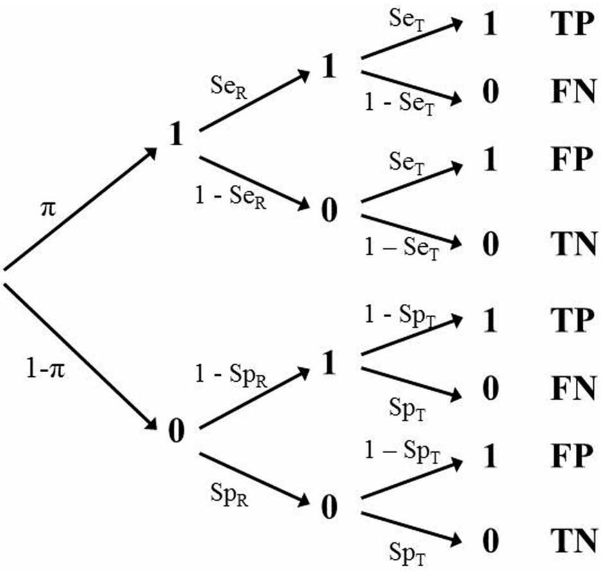
MODEL 2: THE REFERENCE IS IMPERFECT

Sometimes the reference employed in assessment studies is not a real GS (the sensitivity and/or specificity of R, denoted by Se_{R} and Sp_{R}, are less than 1). In psychology, categorical diagnostic assessment is often performed using an in-depth interview where the DSM criteria (American Psychiatric Association, 2002) are checked, combined with complementary tests. This is usually considered an almost perfect reference (a GS, in practical terms). However, in some studies assessing the diagnostic capability of binary classification tools, the diagnosis employed as the reference is not done like this. Instead, it is done with another psychometric test with relatively good properties, but with sub optimal reliability. For example, to assess the accuracy of SCOFF (*Sick*-*Control*-*One*-*Fat*-*Food*; Morgan et al., 1999; Hill et al., 2010), the *Eating Attitudes Test* (*EAT*; Garner et al., 1982) is sometimes used as the reference. In the "short" version of this test (26 items scored 0–3) a cut-off of $X = 20$ is often used for classification (e.g., Berger et al., 2011). Although errors may occur in a diagnostic interview its reliability is undoubtedly higher than that of a test such as *EAT*.

The main difference between model 1 and model 2 is that in model 1 the observed prevalence of the study, $(TP + FN)/M$, is the true empirical prevalence. By contrast, in the scenario implied in model 2 the observed prevalence does not correspond to the true empirical prevalence in the study. In the TP and FN cells there is an unknown fraction of observations that are not real 1s, and/or in the FP and TN cells there is an unknown fraction of observations that are not real zeros.

Consequently the tree model must incorporate new features, represented by new parameters. Specifically, it must include two values of sensitivity (Se_{R}, Se_{T}) and two values of specificity (Sp_{R}, Sp_{T}). This means that some observations that are categorized as TP should actually have been coded as FP, since R had misclassified them as 1 when they were 0. The same is true in the remaining cells. Each of the four observed frequencies are composed of a genuine part of observations, plus a portion that have been counted in the wrong cell, because they were incorrectly classified by R (as 1, when they were a 0, or vice versa; **Figure 3**).

When referring to a single study, this model includes five parameters (π, Se_{T}, Sp_{T}, Se_{R}, Sp_{R}), but there are only three degrees of freedom available. The model cannot be identified or tested. However, if there are several studies with independent estimates then it is possible to properly estimate the parameters.



$$p_{TP} = \pi \cdot Se_R \cdot Se_T + (1-\pi)\cdot(1-Sp_R)\cdot(1-Sp_T)$$

$$p_{FN} = \pi \cdot Se_R \cdot (1-Se_T) + (1-\pi)\cdot(1-Sp_R)\cdot Sp_T$$

$$p_{FP} = \pi \cdot (1-Se_R)\cdot Se_T + (1-\pi)\cdot Sp_R \cdot (1-Sp_T)$$

$$p_{TN} = \pi \cdot (1-Se_R)\cdot(1-Se_T) + (1-\pi)\cdot Sp_R \cdot Sp_T$$

FIGURE 3 | Tree diagram of Model 2 (IR) and the corresponding equations.

As in the previous model, we assume a set of k independent studies, each with a different sampling model and therefore with a different empirical prevalence, π_i. The number of independent parameters to be estimated is $k + 4$. The number of degrees of freedom is $2k - 4$. The minimum number of studies required to fit and test this model is 3, because the parameters are three values of π, plus the two sensitivities and two specificities. The degrees of freedom would be 2.

Although model 2 is appropriate in many situations, it makes a debatable assumption about the independence between the classifications provided by R and T within each category. Thus, it assumes that the probability that the test will yield a positive result when applied to a participant with status "1" remains the same, regardless of whether this case has given a positive or negative result in R. However, a refinement of model 2 that relaxes the assumption of independence between the classifications provided by R and T is not identifiable within the present framework and is not discussed in this paper.

ASSESSING AND SELECTING A MODEL

A model's fit in MTM is usually assessed by the likelihood Ratio Test through the G^2 statistic (Read and Cressie, 1988), asymptotically distributed as χ^2 when the model is true. This statistic allows the decision about whether the model must be rejected or not. But it is possible for more than one model to fit, so the problem of selecting a particular model arises.

If a model has a significantly lower discrepancy it is considered to be a better representation of what is being modeled (Ulrich, 2009). However, it is well-known that more complex models (for example, those with more free parameters) tend to fit better, because they have more flexibility and can capitalize on chance (Pitt et al., 2002). The different levels of complexity must be considered when comparing models. In our context GS and IR are nested models (GS is a particular case of IR, with $Se_R = Sp_R = 1$). When both models fit the data well the difference between their G^2 statistics is asymptotically distributed as χ^2 with two degrees of freedom. So, the incremental fitting can also be tested.

Several criteria and indices have been proposed that essentially reflect the trade-off between the predictive accuracy of the model and the model's complexity. The Akaike (1974) and Bayesian (Schwartz, 1978) criteria have been criticized because they do not fully account for all relevant dimensions of complexity. However, an interesting alternative is the Minimum Description Length criterion, which has recently been proposed for selecting MTM models (Wu et al., 2010). When two models have the same fit the MDL selects the less complex, as the Akaike and Bayesian criteria do. But the MDL also takes into account other dimensions of the models' complexity, such as their functional form. This criterion as implemented in MultiTree (Moshagen, 2010) will be employed in the examples below.

For both GS and IR models, when fitted as in the two examples below (with four independent studies), the rank of the Jacobian matrix is not lower than the number of parameters estimated. Thus, the models are locally identifiable. However, as is shown in the examples, model 2 is always associated with two different sets of estimates, as it shows a problem of global identifiability. When IR models are fitted in repeated occasions on the same data MultiTree provides alternatively two sets of parameters that yield the same (and minimum) value for G^2. Thus, using the data in the two examples below, different runs can produce any of the two sets of parameters in **Table 1**. In each example the specificity in set 2 is the complementary of the sensitivity of set 1 and the sensitivity in set 2 is the complementary of the specificity of set 1. This happens both in the reference and the test.

It could be argued that there is no means to judge which set of estimates is correct. Although in some special cases could be difficult, in most cases it is a matter of common-sense. Both the reference and the test are chosen for the study because they are well-known to be effective. The purpose of the study is to obtain an accurate estimation of the efficacy of the test (and the experience suggests that it works better than tossing a coin). As one of the two sets of parameters involves a less than random performance, the other must be chosen. Even worst, the values in set 2 reflect accurate but perverse tools. They reflect a test which classifies the vast majority of targets as normal and vice versa. So, one of the sets is congruent with the estimates provided by the individual studies while the other is incongruent with them (see the examples below). Nevertheless, global identifiability can be achieved by imposing parametric order constraints (see Knapp and Batchelder, 2004). If when running with MultiTree both examples of **Table 1** it is imposed the constraint that the four parameters reflecting accuracy (Se_T, Sp_T, Se_R, Sp_R) are higher than 0.50 then the correct solution is always reached. However, there could be other types of tests or measurement contexts in which the choice is not as clear as in our examples. Thus, it is possible that only three of those parameters are higher than 0.50. For example, sometimes the target status is difficult to detect by the reference (Se_R around 0.50) but the study is still worthy because specificity is very high. In any study of this type the two solutions must be obtained and compared. It is possible that in none of the two solutions the four parameters are simultaneously higher than 0.50; in those cases other models of measurement are probably needed.

Table 1 | Parameters estimated in the two solutions provided by Multitree for Model 2, IR, in the two examples with real data.

		Se_R	Sp_R	Se_T	Sp_T	G^2
AUDIT	Set 1	0.996	1.00	0.637	0.960	13.985
	Set 2	0.000	0.004	0.040	0.363	
MMSE	Set 1	0.876	1.00	0.864	0.872	12.136
	Set 2	0.000	0.124	0.128	0.136	

TWO EXAMPLES WITH REAL DATA

Applying the procedure requires that the k studies included are homogeneous. This is not a problem in simulation studies, but with real data it is impossible to have studies that are exact replicates. Real data differ in such characteristics as the type of professional who manages R and T, the context in which it is applied, or the language version employed in the test. However, within certain boundaries the studies can and should be homogeneous, enabling the interpretation and recognition of the estimated parameters in all of them. The main objective here—to determine whether or not the reference is a GS—can be addressed with a level of homogeneity that need not be completely strict. However, as a consequence of not using exact replicates some additional misfit must be expected. This is the rationale for preferring among the conventional alpha levels the more liberal for our purposes (0.01 instead of 0.05) when testing the absolute fit of the models. However, power analyses will be performed for every test.

THE TEST AUDIT AND SELF-REPORT OF DRINKING AS A REFERENCE

We used four independent studies to assess the accuracy for the classification of the *Alcohol Use Disorder Identification Test* (*AUDIT*; Babor et al., 2001). These studies shared the same specific target population, the elderly, and used the same reference for the classification. The reference is an objective (although self-reported) amount of alcohol consumption of at least 14 drinks per week. The test is employed with a cut-off value of $X \geq 8$ as a rule for the binary classification in males. The top panel of **Table 2** identifies the studies. It also shows the raw data, the sensitivity, the specificity, and the observed prevalence in each study.

The two top rows of **Table 3** show some of the results provided by MultiTree (Moshagen, 2010) when fitting the two models. There are several reasons for choosing model 1 as the one which better describes the behavior of the reference and the test in this example. Firstly, the goodness-of-fit statistic for the GS model shows an acceptable value ($p > 0.01$). When two more parameters are included (model 2) the fit does not improve significantly (the statistic is virtually equal, but with two more parameters). Secondly, in model 2, the values for Se_{R} and Sp_{R} are both close to 1 (rounding to the third decimal gives 0.996 and 1.0, respectively), the values representing an optimal reference or GS. Thirdly, the criterion employed for model selection, MDL, gives a smaller value for the GS model than for the IR model (see the last column in **Table 3**).

We have also performed a power analysis (*post-hoc* analysis; Faul et al., 2007) of the test for the incremental fit, as follows. In the nested model (GS) the parameters describing the reference are $Se_{\mathrm{R}} = Sp_{\mathrm{R}} = 1$. We have established that for the test being convincing it must have enough power to detect a small-medium effect size in the sense of Cohen (1988; $w = 0.10$ is considered small and $w = 0.30$ is considered medium). In this case the small effect size is obtained (approximately) by setting in the alternative model the values $Se_{\mathrm{R}} = Sp_{\mathrm{R}} = 0.95$ (the effect size is $w = 0.097$). With alpha 0.05 the power is 0.707. The small-medium effect size is obtained (approximately) by setting the values $Se_{\mathrm{R}} = Sp_{\mathrm{R}} = 0.80$ (the effect size is $w = 0.20081$). With alpha 0.05 the power is 0.9997. When alpha is set at 0.01 the power values for the same effect size values are 0.475 (for small w) and 0.998 (for small-medium w).

In summary, the model chosen was model 1, which implies that the reference employed (self-report of at least 14 drinks per week) works most probably as a GS. The best estimates of the sensitivity and specificity of the test are the values obtained using model 1 ($Se_{\mathrm{T}} = 0.637$; $Sp_{\mathrm{T}} = 0.960$). The results of these four studies suggest that whereas the AUDIT almost never classifies a normal behavior as a disorder, (about 4%), it misses a considerable number of alcohol abuse problems in the elderly (slightly more than one-third).

The set of parameters in **Table 3** is the one obtained after imposing the parameter constraints described above (Se_{T}, Sp_{T}, Se_{R}, $Sp_{\mathrm{R}} > 0.50$). Common-sense also advises the same conclusion, as examining the calculated sensitivities and specificities of the studies (**Table 2**) it is clear that only this solution is congruent with the meaning of the parameters.

Table 2 | Raw frequencies of the primary studies included in the two examples with real data.

Test	Study	*TP*	*FN*	*FP*	*TN*	*Se*	*Sp*	Prev.
AUDIT	Bradley et al., 1998	58	47	6	150	0.552	0.962	0.402
	Philpot et al., 2003	13	4	7	104	0.765	0.937	0.133
	Reid et al., 2003, (sample 1)	22	3	6	148	0.880	0.961	0.140
	Reid et al., 2003, (sample 2)	7	3	8	243	0.700	0.968	0.038
MMSE	Brayne and Calloway, 1989	24	5	31	205	0.828	0.869	0.109
	Brodaty et al., 2002	66	16	48	153	0.805	0.761	0.290
	Clarke et al., 1991	137	17	28	122	0.890	0.813	0.507
	Cullen et al., 2005	40	4	138	933	0.909	0.871	0.039

THE TEST MMSE AND THE CAMDEX AS A REFERENCE

We also used four independent studies that provide data allowing the assessment of the accuracy for the classification of the *Mini Mental State Examination* (*MMSE*; Folstein et al., 1975). This is a test for detecting unspecific dementia, although it is often used to detect the early stages of Alzheimer's disease. The test is employed with a rule for classification based on a cut-off value of $X = 24$. The reference employed in the primary studies included the *Cambridge Mental Disorders of the Elderly Examination* (*CAMDEX*; Roth et al., 1986). The bottom panel of **Table 2** identifies the studies and shows the raw data, the sensitivity, the specificity, and the observed prevalence in each study.

Table 3 (two bottom rows) shows the results provided by MultiTree. There are several reasons for choosing here model 2 as the one that better describes the behavior of the reference and the test. Firstly, the goodness-of-fit for model 1 shows a significant deviation between the empirical and predicted frequencies ($p < 0.01$), but model 2 does not ($p > 0.01$). Secondly, although the value for Sp_{R} in model 2 is virtually 1 (the optimal classification of "normals"), the value of Se_{R} is far from its upper boundary. In this particular example, the parameters of model 2 indicate that the reference has virtually perfect specificity, whereas

Table 3 | Parameter estimates, goodness-of-fit and Minimum Description Length of the two models for the AUDIT and the MMSE data.

Test	Model	Reference				Test				Goodness of fit			C_{FIA}	*MDL*
		Se_{R}		Sp_{R}		Se_{T}		Sp_{T}		G^2	*df*	*p*		
		Estim	*SE*	Estim	*SE*	Estim	*SE*	Estim	*SE*					
AUDIT	1—GS	–		–		0.637	0.038	0.960	0.008	13.99	6	0.030	17.9	574.8
	2—IR	0.996	0.054	1.00	0.011	0.637	0.044	0.960	0.011	13.99	4	0.007	20.1	577.0
MMSE	1—GS	–		–		0.864	0.020	0.852	0.009	21.14	6	0.002	20.1	1495.2
	2—IR	0.876	0.040	1.00	0.004	0.864	0.023	0.872	0.011	12.14	4	0.016	23.0	1493.6

the sensitivity is suboptimal ($Se_R = 0.876$). Assuming this value for the sensitivity of the reference, the estimated parameters for the MMSE are $Se_T = 0.864$ and $Sp_T = 0.872$, respectively. Thirdly, the criterion employed for model selection, MDL, provides a smaller value for the IR model than for the GS model (**Table 3**).

A power analysis of the test for the incremental fit test similar to that of the AUDIT yielded the following results. In this case, to achieve a small effect size ($w = 0.099$) the values in the alternative model should be set as $Se_R = Sp_R = 0.975$. With alpha 0.05 the power is 0.982. The small-medium effect size is obtained (approximately) by setting the values $Se_R = Sp_R = 0.91$ (the effect size is $w = 0.20313$). With alpha 0.05 the power is 1 when rounded to three decimals. When alpha is set at 0.01 the power values for the same effect size values are 0.932 (for small w) and 1 (for small-medium w).

As with AUDIT, common sense advises choosing the set 1 in **Table 1** rather than set 2, as the only one that is congruent with the meaning of the parameters. The same set is the obtained after imposing the same constraints as in the AUDIT example. In summary, the model chosen for the MMSE was model 2, which implies that the reference employed in this set of studies (CAMDEX) is most probably an IR. Then, if the accuracy of the MMSE is assessed without acknowledging that the reference has suboptimal reliability, then the estimate of the accuracy of the test is biased.

The results of the example with the MMSE test can give the impression that in the end the difference is small and not worth the effort to study and assess the accuracy of the reference. Consider a numerical example with more dramatic results. Suppose the test X has $Se_X = 0.90$ and $Sp_X = 0.85$ and the accuracy is evaluated using a gold standard reference. If there are 200 target cases in the study and 800 normal cases (observed prevalence = empirical prevalence = 0.20) the expected frequencies are 180 (TP), 20 (FN), 120 (FP), and 680 (TN). However, if in the same study the reference is imperfect and has $Se_R = 0.90$ and $Sp_R = 0.90$ then the expected frequencies are 174 (TP), 86 (FN), 126 (FP), and 614 (TN). As a consequence, if the imperfection of the reference is not acknowledged the expected estimates of the sensitivity and specificity are 0.669 [calculated as $174/(174 + 86)$] and 0.830 [calculated as $[614/(614 + 126)]$], respectively, instead of the actual 0.90 and 0.85 values. Furthermore, although the empirical prevalence is 0.20 the observed prevalence is now 0.26 [calculated as $(614 + 126)/1000$].

In short, if researchers do not recognize the imperfection of the reference they will not be aware that dozens of cases have been misclassified in his/her data, and will report underestimations of the accuracy of the test. Furthermore, this could lead to unsatisfactory choice when several tests are available as candidates for the same diagnostic purpose, as decision is based on comparisons of their accuracies. An apparently better test would be preferred only because its diagnostic accuracy has been assessed against a gold standard reference. Other equally or even more suitable tests would be eliminated because they have been validated using imperfect references. The choice of a test for the screening process will be flawed.

DISCUSSION

The reliability of the reference must be evaluated and taken into account when assessing the accuracy of tools for binary classification in screening processes (Valenstein, 1990). Whilst in most statistical models it is assumed that the reference is perfectly reliable (GS), in psychology the references usually employed are suboptimal. In medicine, the references are sometimes objective states that can be checked with almost perfect accuracy. But in psychology we often lack such references. Construct validity is an enduring concern for researchers in psychology (Cook and Campbell, 1979; Messik, 1989, 1995), because we know that the references are almost always suboptimal. In some cases the accuracy of the reference is so high that it can be taken as a GS without having any relevant impact on the results of the estimates. However, as a general rule their status should be assessed.

Sometimes it is acknowledged that the reference used in studies that assess the accuracy of screening tools may be imperfect. Procedures have been proposed to account for this suboptimal reliability in estimating the accuracy of the test (Rutjes et al., 2007; Reitsma et al., 2009; Trikalinos and Balion, 2012). They include, for example, combining several references to yield a single, better criterion. However, usually in these procedures it is assumed that the researchers already know whether their reference is GS or IR.

We have proposed a procedure which satisfies the need for distinguishing between two different assessment scenarios: perfect reference (GS) vs. imperfect reference (IR). Nested multinomial tree models built with the parameters that define those two models are fit. The two examples described show that sometimes it is better to assume that the reference is a GS (knowing that its reliability is not perfect, but virtually optimal), but at other times the suboptimality of the reference must be acknowledged. The identification of a reference as a gold standard or as an imperfect reference allows a better evidence-based choice of the better test available for a specific screening process.

An alternative to the approach taken here could be useful when there are no independent and homogeneous studies available, but there is a single study with a large sample. This alternative consists of a random partition of the sample into multiple segments. Although problems may arise due to the sub-samples are extracted from exactly the same large sample, this alternative should be evaluated in future simulation studies.

Psychometric meta-analysis (Vacha-Haase, 1998; Hunter and Schmidt, 2004; Rodriguez and Maeda, 2006; Botella et al., 2010) provides combined estimates of the psychometric properties of the data obtained with a given test. The procedure outlined in this paper can be applied to refine meta-analytic estimates of the accuracy of tests employed for binary classifications. Such meta-analyses integrate primary studies that assess the accuracy of binary classifications performed on a specific test (e.g., Botella et al., 2013). If the primary studies included in a meta-analysis have been carried out using an IR, the combined estimate of the accuracy will be incorrect, unless the imperfection of the reference is detected and assessed. The present procedure allows the detection of such suboptimal performance.

A limitation of the procedure is that it is not yet capable of managing situations where the classifications provided by the test are not conditionally independent of the classification provided by the reference.

ACKNOWLEDGMENTS

This research was supported by grant PSI2009-12071 (Ministerio de Ciencia e Innovación, Spain). Thanks are due to reviewer Edgar Erdfelder, who directly suggested reparameterization as a way to reach global identifiability in model 2. Thanks are also due to Miguel Angel García-Pérez for several comments and suggestions to a previous version of the manuscript, and specially the alternative of a random partition of a large sample in several subsamples.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

American Psychiatric Association. (2002). *Diagnostic and Statistical Manual of Mental Disorders, DSM-IV-TR, 4th Edn.* Washington: American Psychiatric Association.

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., and Monteiro, M. G. (2001). *The Alcohol Use Disorders Identification Test: Guidelines for Use in Primary Care, 2nd Edn.* Geneva, Switzerland: World Health Organization.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychol. Assess.* 10, 331–344. doi: 10.1037/1040-3590.10.4.331

Batchelder, W. H., and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* 6, 57–86. doi: 10.3758/BF03210812

Berger, U., Wick, K., Hölling, H., Schlack, R., Bormann, B., Brix, C., et al. (2011). Screening riskanten Essverhaltens bei 12-jährigen Mädchen und Jungen: psychometrischer Vergleich der deutschsprachigen Versionen von SCOFF und EAT-26. *Psychother. Psych. Med. Psychol.* 61, 311–318. doi: 10.1055/s-0031-1271786

Botella, J., Sepúlveda, A. R., Huang, H., and Gambara, H. (2013). A meta-analysis of the diagnostic accuracy of the SCOFF. *Span. J. Psychol.* (accepted).

Botella, J., Suero, M., and Gambara, H. (2010). Psychometric inferences from meta-analysis of reliability and internal consistency coefficients. *Psychol. Methods* 15, 386–397. doi: 10.1037/a0019626

Bradley, K. A., Bush, K. R., McDonell, M. B., Malone, T., and Fihn, S. D. (1998). Screening for problem drinking: comparison of CAGE and AUDIT. *J. Gen. Intern. Med.* 13, 379–389. doi: 10.1046/j.1525-1497.1998.00118.x

Brayne, C., and Calloway, P. (1989). An epidemiological study of dementia in a rural population of elderly women. *Br. J. Psychiatry* 155, 214–219. doi: 10.1192/bjp.155.2.214

Brodaty, H., Pond, D., Kemp, N. M., Luscombe, G., Harding, L., Berman, K., et al. (2002). The GPCOG: a new screening test for dementia designed for general practice. *J. Am. Geriatr. Soc.* 50, 530–534. doi: 10.1046/j.1532-5415.2002.50122.x

Clarke, M., Jagger, C., Anderson, J., Battcock, T., Kelly, F., and Stern, M. C. (1991). The prevalence of dementia in a total population: a comparison of two screening instruments. *Age Ageing* 20, 396–403. doi: 10.1093/ageing/20.6.396

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Erlbaum.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation Design and Analysis Issues for the Field Settings.* Boston, MA: Houghton Mifflin.

Cullen, B., Fahy, S., Cunningham, C. J., Coen, R. F., Bruce, I., Greene, E., et al. (2005). Screening for dementia in an Irish community sample using MMSE: a comparison of norm-adjusted versus fixed cut-points. *Int. J. Geriatr. Psychiatry* 20, 371–376. doi: 10.1002/gps.1291

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models: a review of the literature. *Z. Psychol./J. Psychol.* 217, 108–124. doi: 10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). 'Mini-Mental State'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

García-Pérez, M. A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *Br. J. Math. Stat. Psychol.* 47, 247–282. doi: 10.1111/j.2044-8317.1994.tb01037.x

Garner, D. M., Olmsted, M. P., Bohr, Y., and Garfinkel, P. (1982). The eating attitudes test: psychometric features and clinical correlates. *Psychol. Med.* 12, 871–878. doi: 10.1017/S0033291700049163

Hill, L. S., Reid, F., Morgan, J. F., and Lacey, J. H. (2010). SCOFF, the development of an eating disorder screening questionnaire. *Int. J. Eat. Disord.* 43, 344–351. doi: 10.1002/eat.20679

Hu, X., and Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika* 59, 21–47. doi: 10.1007/BF02294263

Hunter, J. E., and Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, 2nd Edn.* Thousand Oaks, CA: Sage.

Knapp, B. R., and Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *J. Math. Psychol.* 48, 215–229. doi: 10.1016/j.jmp.2004.03.002

Messik, S. (1989). "Validity," in *Educational Measurement*, ed R. L. Linn (New York, NY: Macmillan), 13–103.

Messik, S. (1995). Validity of psychological assessment. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Morgan, J. F., Reid, F., and Lacey, J. H. (1999). The SCOFF questionnaire: assessment of a new screening tool for eating disorders. *Br. Med. J.* 319, 1467–1468. doi: 10.1136/bmj.319.7223.1467

Moshagen, M. (2010). MultiTree: a computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* 42, 42–54. doi: 10.3758/BRM.42.1.42

Moshagen, M., and Hilbig, B. E. (2011). Methodological notes on model comparisons and strategy classification: a falsificationist proposition. *Judgm. Decis. Mak.* 6, 814–820.

Philpot, M., Pearson, N., Petratou, V., Dayanandan, R., Silverman, M., and Marshall, J. (2003). Screening for problem drinking in older people referred to a mental health service: a comparison of CAGE and AUDIT. *Aging Ment. Health* 7, 171–175. doi: 10.1080/1360786031000101120

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychol. Rev.* 109, 472–491. doi: 10.1037/0033-295X.109.3.472

Read, T. R. C., and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New York, NY: Springer. doi: 10.1007/978-1-4612-4578-0

Reid, M. C., Tinetti, M. E., O'Connor, P. G., Kosten, T. R., and Concato, J. (2003). Measuring alcohol consumption among older adults: a comparison of available methods. *Am. J. Addict.* 12, 211–219. doi: 10.1111/j.1521-0391.2003.tb00649.x

Reitsma, J. B., Rutjes, A. W., Khan, K. S., Coomarasamy, A., and Bossuyt, P. M. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J. Clin. Epidemiol.* 62, 797–806. doi: 10.1016/j.jclinepi.2009.02.005

Rodriguez, M., and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychol. Methods* 11, 306–322. doi: 10.1037/1082-989X.11.3.306

Roth, M., Tym, E., Mountjoy, C. Q., Huppert, F. A., Hendrie, H., Verma, S., et al. (1986). CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *Br. J. Psychiatry* 149, 698–709. doi: 10.1192/bjp.149.6.698

Rutjes, A. W. S., Reitsma, J. B., Coomarasamy, A., Khan, K. S., and Bossuyt, P. M. M. (2007). Evaluation of diagnostic tests when there is no gold standard: a review of methods. *Health Technol. Assess.* 11, 9–51.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Trikalinos, T. A., and Balion, C. M. (2012). Chapter 9: options for summarizing medical test performance in the absence of a 'gold standard'. *J. Gen. Intern. Med.* 27(Suppl. 1), 67–75. doi: 10.1007/s11606-012-2031-7

Ulrich, R. (2009). "Uncovering unobservable cognitive mechanisms: the contribution of mathematical models," in *Neuroimaging of Human Memory: Linking Cognitive Processes to Neural Systems*, eds F. Rösler, C. Ranganath, B. Röder, and R. H. Kluwe (New York, NY: Oxford University Press), 25–41. doi: 10.1093/acprof:oso/9780199217298.003.0003

Vacha-Haase, T. (1998). Reliability generalization: exploring variance in measurement error affecting score reliability across studies. *Educ. Psychol. Meas.* 58, 6–20. doi: 10.1177/0013164498058001002

Valenstein, P. N. (1990). Evaluating diagnostic tests with imperfect standards. *Am. J. Clin. Pathol.* 93, 252–258.

Wu, H., Myung, J. I., and Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychon. Bull. Rev.* 17, 275–286. doi: 10.3758/PBR.17.3.275

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 July 2013; accepted: 12 September 2013; published online: 03 October 2013.

Citation: Botella J, Huang H and Suero M (2013) Multinomial tree models for assessing the status of the reference in studies of the accuracy of tools for binary classification. Front. Psychol. ***4****:694. doi: 10.3389/fpsyg.2013.00694*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.