МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2016
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2016
Proceedings of the Forty Fifth Spring Conference
of the Union of Bulgarian Mathematicians
Pleven, April 6–10, 2016

ONE METHOD TO CHECK THE POPULATION HOMOGENEITY OF A TEST*

Dimiter Tsvetkov, Lyubomir Hristov, Ralitsa Angelova-Slavova

In this paper it is considered a method for checking the population homogeneity of a given psychological or didactical test with respect to the way of its accepting from different populations. For this purpose we look for a statistical association between the a priori known populations and posterior estimated clusters on the base of experimental data. The lack of such statistically significant association is considered as an evidence for the absence of a discrimination and therefore as a strong indicator for homogeneity. The posterior clusters are searched by means of the Generalized Partial Credit Model from the Item Response Theory.

The basic probability model for the typical tests. The psychological measurement represents assigning of a numerical or some other sign characteristic to the observed person [1, 12]. In the modern test theory – Item Response Theory [2, 3, 9, 10, 11, 17, 18, 19] (IRT) the ultimate aim of the measurement is to quantify some specific latent variable θ (ability). Actually the test result v (the answers set) presents just some result indicator which must be transformed to usual measurement. Thus the measurement itself can be considered as composed from two stages. At the first stage we postulate a stochastic relation between the ability θ and the indicator v. At the second stage we have to find some method to obtain statistical estimation $\hat{\theta}$ on the base of v. The typical test consists of a finite number I of test items. The result from the item i execution is rendered by successive markers $0, 1, \ldots, M_i$ (the integers from 0 to M_i.). The probability model for the test item i consists of the assumption for the presence of the probabilities $P_{im}(\theta)$ (Item Category Response Function) for obtaining a result with a marker m under a performance of an individual of ability θ. These probabilities naturally depend on a set of parameters $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots)$. The set of all item parameters is denoted by $\eta = (\xi_1, \xi_2, \ldots, \xi_I)$. In this text we stick to the traditional assumption of local stochastic independence of test results for separate items.

The result indicator can be written as a matrix $v = (v_{im})$, $1 \leq i \leq I$, $0 \leq m \leq M_i$, where $v_{im} = 1$ if for item i is pointed marker m and $\nu_{im} = 0$ otherwise. Also it is convenient for a datum to be presented as vector $u = (u_i)$ in which each coordinate points the corresponding marker. Without loss of generality we assume a normal $N(\mu, \sigma^2)$ population distribution of the ability θ with a density $\varphi(\theta)$ whose parameters are considered as an overall model ones.

* **2010 Mathematics Subject Classification:** 97K80, 62P25.
Key words: Population Homogeneity, Item Response Theory.

Under the above assumptions, the probability of a result indicator v is given by the formula (see e.g. [18, 19])

$$f(v|\theta,\eta)=\prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta,\xi_i)\right)^{v_{im}}$$

conditionally to θ which defines a discrete probability space over the set of the result indicators. The joint distribution of the random variables v and θ has density

$$f(v,\theta|\eta)=f(v|\theta,\eta)\,\varphi(\theta|\eta)=\left(\prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta,\xi_i)\right)^{v_{im}}\right)\varphi(\theta|\eta) \tag{1}$$

and therefore, for any indicator set V and for any numerical interval Δ the probability of the basic events $(v\in V)\,\&\,(\theta\in\Delta)$ is given by the formula

$$\Pr\left((v\in V)\,\&\,(\theta\in\Delta)\right)=\sum_{v\in V}\int_{\Delta}f(v,\theta|\eta)\,d\theta.$$

Further all the integrals are taken from $-\infty$ to ∞ and the notation η stands for the overall model parameters set. The joint distribution (1) we call a basic probability model because it describes the relation between the aim of the measurement θ and its indicator v.

The marginal distribution $f(\nu|\eta)$ of v can be obtained after elimination of θ from the joint distribution

$$f(v|\eta)=\int f(v,\theta|\eta)\,d\theta=\int\left(\prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta,\xi_i)\right)^{v_{im}}\right)\varphi(\theta|\eta)\,d\theta. \tag{2}$$

The generation of an artificial data for (1) is made through the scheme $\theta_{(j)}\sim\varphi(\theta|\eta)$, $v_{(j)}\sim f\left(v|\theta_{(j)},\eta\right)$. In this way $(v_{(j)},\theta_{(j)})$, $1\le j\le J$, forms the complete data for (1) and $(v_{(j)})$ becomes an *iid* sample of result indicators. Let we are given an *iid* sample $v=(v_{(j)})$. Then according to (2) its likelihood function has the form

$$L(v|\eta)=\prod_{j=1}^{J}f\left(v_{(j)}|\eta\right)=\prod_{j=1}^{J}\int f\left(v_{(j)},\theta|\eta\right)d\theta$$

$$=\prod_{j=1}^{J}\int\left(\prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta,\xi_i)\right)^{v_{(j)im}}\right)\varphi(\theta|\eta)\,d\theta$$

from which for the log-likelihood we find

$$l(v|\eta)=\ln L(v|\eta)=\sum_{j=1}^{J}\ln f\left(v_{(j)}|\eta\right) \tag{3}$$

$$=\sum_{j=1}^{J}\ln\int\left(\prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta,\xi_i)\right)^{v_{(j)im}}\right)\varphi(\theta|\eta)\,d\theta.$$

The estimation of the parameters η under the presentation (3) is done by the popular EM-algorithm.

The EM-algorithm. The EM-algorithm [5] is a very basic and common iterative method for obtaining a maximum likelihood estimation (MLE) for the parameters η. As

a stop criteria one can use a sufficiently small change in $\ln L(v|\eta)$ value or sufficiently small absolute or relative change in the successive values of η. The EM-algorithm has various favorable properties from the numerical point of view – it is stable and convergent in the typical case. In the common case however it may converge to a saddle point [13].

The M-step of the algorithm requires solving of the multivariable optimization task (see e.g. [6, 15, 16]). In our case the optimization problem appears relatively simple because it is performed separately item by item. This fact makes the use of the Newton-Raphson method very convenient and effectual. Here it is possible to use also various quasi Newton methods as BFGS and also some none typical methods as genetic algorithms. However the success of the iterative methods depends heavily on the choice of the initial value.

The presence of an integrals in the analytical terms requires a numerical integration of certain integrals with a weight function the standard normal density $\varphi(\cdot|\sigma)$ with zero mean and dispersion σ^2. In this paper we follow the traditional approach to use the Gaussian quadrature formulas. This approach has an advantage that the numerical integration can be considered as a replacement of the continuous normal ability distribution by a discrete distribution supported on the quadrature nodes (θ_s) and with probabilities corresponding quadrature weights (w_s).

Polytomous models. Using of the binary test items appears as a sort of a situation in which the possible outcomes are minimal in number (only 2). In this case the forced polarization of the responses has either obvious advantages and disadvantages. For marking the middle (neutral) answer diapason we need at least one more marker [9, 10, 11]. In the practice they are used 4 or 5 outcome markers per test item. A great variety of such polytomous IRT models are presented for example in [9]. Here we pay attention to the so called Generalized Partial Credit Model (GPCM). In GPCM the markers m are whole numbers from to some $M \geq 1$. The probability for result m of a person of ability θ is

$$P_m(\theta) = \frac{\exp\left(\sum\limits_{s=0}^{m} a(\theta - d_s)\right)}{\sum\limits_{\mu=0}^{M_i} \exp\left(\sum\limits_{s=0}^{\mu} a(\theta - d_s)\right)}, \tag{4}$$

where (d_s) are parameters for item localization and the parameter a is associated with the property of an item discrimination power. The parametric uncertainty in (4) is avoided by putting $d_0 = 0$ after which for item i we have

$$P_{i0}(\theta, a_i, b_{i1}, \ldots, b_{iM_i}) = \frac{1}{1 + \sum\limits_{\mu=1}^{M_i} \exp(\mu a_i \theta + b_{i\mu})},$$

$$P_{im}(\theta, a_i, b_{i1}, \ldots, b_{iM_i}) = \frac{\exp(m a_i \theta + b_{im})}{1 + \sum\limits_{\mu=1}^{M_i} \exp(\mu a_i \theta + b_{i\mu})},$$

where $b_{im} = -a_i(d_{i1} + \cdots + d_{im})$. Now for the parameters ξ_i of test item i we have

$\xi_i = (a_i, b_{i1}, \ldots, b_{iM_i})$. The basic probability model turns into the form

$$f(v, \theta|\eta) = f(v|\theta, \eta)\,\varphi(\theta) = \prod_{i=1}^{I}\prod_{m=0}^{M_i}\left(P_{im}(\theta, \xi_i)\right)^{v_{im}}\varphi(\theta),$$

where $\eta = (\xi_1, \ldots, \xi_I)$ is the overall model parameter set. The population distribution of the ability θ is assumed to be $N(0,1)$ by which is avoided the uncertainty in the parameter a.

Models with latent classes. In the models with latent classes (see e.g. [4, 7, 8]) it is assumed that the population is separated in some parts numbered with k, $1 \le k \le K$. Let α_k defines the probability for a random choice from the population k. This choice is associated with a random variable κ accepting values k with probabilities $f(\kappa = k) = \alpha_k$. The joint distribution of (ν, θ, k) has a density

$$f(v, \theta, k|\eta) = f(v|\theta, \eta_k)\,\varphi(\theta|\eta_k)\,\alpha_k, \quad \left(\sum_k\sum_v \int f(v, \theta, k|\eta_k)\,d\theta = 1\right) \tag{5}$$

where $f(v|\theta, \eta_k)\varphi(\theta|\eta_k)$ sets some basic probability model for population k with parameters η_k and $\eta = (\eta_k, \alpha_k)$ defines the aggregation of all parameters of (5). The right-hand side of (5) defines a mixture of distributions. Here we have two latent variables θ and k therefore the complete data of a sample of size J has the form $(v_{(j)}, \theta_{(j)}, k_{(j)})$, $1 \le j \le J$. Here the values $\theta_{(j)}$ and $k_{(j)}$ are not observable in the experiment. The generation of an artificial complete data for (5) is made through the following schema

$$k_{(j)} \sim (\alpha_k), \quad \theta_{(j)} \sim \varphi\left(\theta|\eta_{k_{(j)}}\right), \quad v_{(j)} \sim f\left(v|\theta_{(j)}, \eta_{k_{(j)}}\right).$$

In this case $data = (v_{(j)})$ becomes an *iid* sample from result indicators for the model (5). Typically the populations models are assumed of the same type and the separate classes differ only in the parametric set. For the logarithmic likelihood function we have

$$\ln f(data|\eta) = \sum_{j=1}^{J}\ln\sum_k \int f\left(v_{(j)}, \theta, k|\eta\right)d\theta = \sum_{j=1}^{J}\ln\sum_k \alpha_k \int f\left(v_{(j)}|\theta, \eta_k\right)\varphi(\theta|\eta_k)\,d\theta.$$

Respectively in the EM-algorithm it holds

$$\mathbf{Q}\left(\eta|\eta^{(t)}\right) = \sum_{j=1}^{J}\sum_k \int \ln f\left(v_{(j)}, \theta, k|\eta\right) f\left(\theta, k|v_{(j)}, \eta^{(t)}\right)d\theta, \; f(\theta, k|v, \eta)$$

$$= \frac{f(v, \theta, k|\eta)}{\sum\limits_k \int f(v, \theta, k|\eta)\,d\theta},$$

which defines the following EM-algorithm – the common scheme for the latent class models.

1	Initialization – choice of initial $\eta^{(0)} = (\eta_k^{(0)}, \alpha_k^{(0)})$
2	E-step. Under known $\eta^{(t)} = \left(\eta_k^{(t)}, \alpha_k^{(t)}\right)$ consider $f\left(\theta, k|v_{(j)}, \eta^{(t)}\right) = \dfrac{f\left(v_{(j)}|\theta, \eta_k^{(t)}\right)\varphi\left(\theta|\eta_k^{(t)}\right)\alpha_k^{(t)}}{\sum\limits_k \alpha_k^{(t)} \int f\left(v_{(j)}|\theta, \eta_k^{(t)}\right)\varphi\left(\theta|\eta_k^{(t)}\right)d\theta}$

3	M-step. Actualize $\eta \to \eta^{(t+1)}$ by the rule $\eta^{(t+1)} = \arg\max_{\eta} \left(\sum_{j=1}^{J} \sum_{k} \int \ln(f(v_{(j)}\|\theta, \eta_k)\varphi(\theta\|\eta_k)\alpha_k) f(\theta, k\|v_{(j)}, \eta^{(t)}) d\theta \right.$ $\left. + [\ln f(\eta\|\tau)] \right)$
4	Stop criteria. Repeat the steps $2 \to 3$ until some stop criteria are met.

The addend $[\ln f(\eta|\tau)]$ takes part only in the case of the presence of the prior distributions for η. The optimization about η_k is made as in the case of a single class model. In the optimization for the parameters α_k it is used the Lagrange multipliers technic from which we get the following actualization formula

$$\alpha_k^{(t+1)} = \frac{1}{J} \sum_{j=1}^{J} \int f\left(\theta, k|v_{(j)}, \eta^{(t)}\right) d\theta$$

In the GPCM case for the stop criteria we use a small enough change in the numerical logarithmic likelihood function

$$\sum_{j=1}^{J} \ln \left(\sum_{k} \alpha_k^{(t)} \sum_{s} w_s \prod_{i=1}^{I} \prod_{m=0}^{M_i} \left(P_{im}\left(x_s, \xi_{ki}^{(t)}\right)\right)^{v_{(j)im}} \right) + \left[\ln f\left(\eta^{(t)}|\tau\right)\right],$$

where the quadrature nodes (θ_s) and weights (w_s) are taken for the Gaussian integration with standard normal distribution density as a weight function and the addend $[\ln f(\eta^{(t)}|\tau)]$ attends again only in the case of the presence of the prior distributions for η.

Let we are given a parameter estimation $\hat{\eta} = (\hat{\eta}_k, \hat{\alpha}_k)$ under some experimental data. Then we are able to classify a person j to population k_j using naturally the highest posterior probability

$$k_j = \arg\max_{k} f\left(k|v_{(j)}, \hat{\eta}\right).$$

Recognizing the prior populations serves as a criterion, to which content the described scheme is efficient (in a certain sense).

Example 1. In this example we consider an illustrative performance of the GPCM classes model for artificial data with $K = 3$ class populations, $I = 20$ test items with $M = 4$ maximal marker per item and with $J = 300$ (100 persons per class population).

Table 1. Classification results for example 1

	posterior class 1	posterior class 2	posterior class 3
population 1	2	95	3
population 2	100	0	0
population 3	0	2	98

This example shows practically perfect posterior recognition. Only 7 person are misclassified from 300 cases.

Example 2. In this example we consider a performance of the GPCM classes model for real experimental data from test about the stress resistance [14] with $K = 2$ populations, $I = 21$ test items with $M = 3$ and a sample with $J = 193$ individuals – males and females.

Table 2. Classification results for example 2

	posterior class 1	posterior class 2
males	64	50
females	41	38

In the last four-cells table we have $\chi^2(1) = 0.338$ with $p = 0.561$ which shows statistical independence between the variable "sex" and variable "stress resistance". This independence points out that the stress resistance test is accepted in the same way by males and females. In other words here is missing a sexual discrimination in the treating of the both sexes.

Conclusions. The great diversity in the IRT models reveals an opportunity to embrace the behavioral uncertainty during the psychological and educational measurements using highly unified approach of the basic probability model and the associated tools.

Using of the models with latent classes offers a rigorous argument to prove the test homogeneity with respect to some biological or social denotation by means of the posterior classification.

The presented method suffers from drawbacks associated with possible unrealistic estimated parameter values. This phenomenon may be covered by using of prior parameter distributions. The authors are thankful to the referee for the useful remarks.

REFERENCES

[1] A. Anastasi. Psychological testing. New York: Macmillan, 1988.

[2] F. Baker. Item response theory: Parameter estimation techniques. New York: M. Dekker, 1992.

[3] F. Baker, S-H. Kim. Item response theory: Parameter estimation techniques. New York: M. Dekker, 2004.

[4] M. Davier, C. Carstensen. Multivariate and mixture distribution Rasch models: Extensions and applications. New York: Springer, 2011.

[5] A. Dempster, N. Laird, D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, No 1 (1977), 1–38.

[6] D. Himmelblau. Applied nonlinear programming. New York: McGraw-Hill, 1972.

[7] L. Hristov, D. Tsvetkov. Parameter Estimation for Finite Mixtures of Partial Credit Models. Internaional Conference on Computer Systems and Technlologies – CompSysTech'2007, Rousse, 2007.

[8] L. Hristov, D. Tsvetkov. Parameter Estimation for Finite Mixtures of Generalized Partial Credit Models. Internaional Conference on Computer Systems and Technlologies CompSysTech'2008, Gabrovo, 2008.

[9] W. Linden, R. Hambleton. Handbook of modern item response theory. New York: Springer, 1997.
[10] F. Lord. Applications of item response theory to practical testing problems. Hillsdale, N. J, Lawrence Erlbaum Associates, 1980.
[11] F. Lord, M. Novick. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley Pub. Co, 1968.
[12] S. Urbina. Essentials of Psychological Testing. Hoboken: John Wiley & Sons, Inc., 2011.
[13] M. Watanabe, K. Yamaguchi. The EM algorithm and related statistical models. New York: Marcel Dekker, 2004.
[14] С. Будева. Личност и адиктивно поведение, дисертация за присъждане на образователна и научна степен доктор, София, 2011.
[15] Б. Боянов. Лекции по числени методи. София: Дарба, 1998.
[16] Б. Сендов, В. Попов. Числени методи. Първа част. София: НИ, 1976.
[17] Е. Стоименова. Измерителни качества на тестовете. София: НБУ, 2000.
[18] Л. Христов. Оценка на параметрите за някои вероятностни модели от теорията на тестовете, дисертация за присъждане на образователна и научна степен доктор, Велико Търново, 2011.
[19] Р. Ангелова-Славова. Приложни аспекти на вероятностни модели от теорията на тестовете, дисертация за присъждане на образователна и научна степен доктор, Велико Търново, 2014.

Dimiter Petkov Tsvetkov
e-mail: dimiter.petkov.tsvetkov@gmail.com
Lyubomir Yanakiev Hristov
e-mail: lyubomir.hristov@gmail.com
Department of "Mathematical Analysis and Applications"
Faculty of Mathematics and Informatics
St. Cyril and St. Methodius University of Veliko Tarnovo
5000 Veliko Tarnovo, Bulgaria

Ralitsa Lubomirova Angelova-Slavova
e-mail: r.angelova.slavova@abv.bg
Vasil Levski National Military University at Veliko Tarnovo
5000 Veliko Tarnovo, Bulgaria

ЕДИН МЕТОД ЗА ПРОВЕРКА НА ХОМОГЕННОСТ НА ТЕСТ

Димитър Петков Цветков, Любомир Янакиев Христов, Ралица Любомирова Ангелова-Славова

В настоящата работа се разглежда метод за проверка на хомогенност на даден психологически или образователен тест относно начина на възприемане от различни популации. За тази цел се търси статистическа асоциация между известните априорни популации и оценени от експериментални данни апостериорни клъстери. Отсъствието на статистически значима асоциация при това съпоставяне се разглежда като обективен индикатор за отсъствие на дискриминация и респективно наличие на еднородност. Апостериорните клъстери се търсят въз основа на обобщения модел на частичния кредит (Generalized Partial Credit Model).