XXI IMEKO World Congress “Measurement in Research and Industry”
August 30 – September 4, 2015, Prague, Czech Republic

ESTIMATING LATENT ABILITY FROM THE NDT RESULTS, WHEN THE TEST ITEMS DIFFICULTIES ARE UNKNOWN BEFOREHAND

Emil Bashkansky [1], *Vladimir Turetsky* [2]

[1] Ind. Eng. & Management Dept., ORT Braude College, Karmiel, Israel, ebashkan@braude.ac.il
[2] Dept. of Applied Mathematics, ORT Braude College, Karmiel, Israel, turetsky1@braude.ac.il

Abstract – A new approach to evaluation of binary test results when checking a one-dimensional ability is proposed. We consider the case where a population of objects is tested by a set of test items having different, but unknown beforehand levels of difficulty, and we need to evaluate both the intrinsic abilities of these objects and the level of difficulty of the test items. We assume that the same scale invariant item response model applies to all members of the tested population of objects under study.

Keywords: binary testing, item response, maximum likelihood

1. INTRODUCTION

In this paper, we deal with the basic and simplest issue of unidimensional ability, when the *test item* performance of the object under test (OUT) can be explained by a single latent ability. OUT can be: program units, electronic components, materials, network connectivity, etc. We consider the case when a qualitatively homogeneous population of OUTs is tested using a set of non-destructive test items having different, but *unknown beforehand* levels of difficulty, and we need to evaluate both the intrinsic abilities of these OUTs and the difficulties of the test items. This set hereinafter will be called the test and it can include any – but must be the same for all OUTs – number of test items. Usually, it is assumed that the test item response is estimated on the binary scale base (pass/fail) and the results of different test items, applied to the same OUT, are conditionally independent (i.e., the response to one test item does not affect the response to another). Homogeneity here means that the same item response model is applied to all members of the population. Even in such a simplified model the matter of effective valuation of test results has not been resolved completely and is still a subject of discussion (see [1] and references therein). In view of this, it seems desirable to develop some unifying/standardized approach for evaluating test results of such tests. We propose an algorithm for test result evaluation applicable to a broad spectrum of engineering tests satisfying the model assumptions described below. The proposed approach combines several already developed methods, allowing building a reasonable numerical scheme for test results evaluation. The developed algorithm is illustrated by a numerical example.

2. TESTING MODEL

Before focusing on the details of the testing model, we would like to make some general considerations. Suppose the studied ability-a is distributed among the tested population of OUT-s according to some probability density function $f(a)$. It may be a discrete distribution, but at the moment, this does not matter, because our aim is to illustrate the general idea. Let d denote the difficulty of the test item in relation to studied ability. We assume that there is some known or supposed function, customary called *item response function* (IRF) - $P(d|a)$, that expresses the probability that the OUT with ability a successfully overcomes the test item of difficulty d. It is natural to assume that the more is d the less is this probability for every given a, i.e. $\frac{\partial P(d|a)}{\partial d} < 0$ Then, the proportion $p(d)$ of OUTs which successfully overcome the test item having difficulty d is expected to be some convolution of $f(a)$ with $P(d|a)$:

$$p(d) = \int P(d|a) f(a) da \qquad (1)$$

Certainly, since $\frac{\partial P(d|a)}{\partial d} < 0$, also $\frac{\partial p(d)}{\partial d} < 0$, which simply means that the more difficult is the test item, the less OUT-s pass it successfully.

It would seem, that in the mathematical sense (1) represents a classic measurement inverse problem (Fredholm integral equation of the first kind), in which one wants to restore the measured value $f(a)$ on the basis of studying the response $p(d)$ of the measuring system given response function $P(d|a)$. However, the problem is that the test items difficulties in our case *are unknown beforehand* as is also $p(d)$ .This circumstance significantly complicates the evaluation problem and its solution.

2.1. Model description

Let us assume that some population of N OUTs is tested by the same test consisting of K test items of unknown beforehand difficulties d_k, $k = 1, 2, ...K$. Every OUT is tested independently, i.e. results of one OUT do not affect the results of other. *A posteriori* the numbering of the test items can be made in order of decreasing frequency of OUTs that successfully passed every test item, hence, it is reasonable to assume that $d_{k+1} \geq d_k$. In total, there exist 2^K possible test results and this resolution does not allow us to reason more than 2^K distinct levels of ability. It is obvious, that the more is K, the better is resolution. Using $"0"$ to indicate failure and $"1"$ to indicate the successful completion of test item, one can present the results of each OUT test as an ordered sequence of length K consisting of zeros and ones. For example, for the test consisting from *K*=3 items, all possible results are presented by their respective *binary codes* as shown below in Table1:

Table 1. Binary codes for *K*=3.

Sequence no.	Test item *1*	Test item *2*	Test item *3*
1	**0**	**0**	**0**
2	**1**	**0**	**0**
3	**0**	**1**	**0**
4	**0**	**0**	**1**
5	**1**	**1**	**0**
6	**1**	**0**	**1**
7	**0**	**1**	**1**
8	**1**	**1**	**1**

2.2. A notational interlude

- $X_{sequence}$ - denotes any notion X relating to a corresponding sequence, namely
    - $p_{sequence}$ - proportion of OUTs whose test results are consistent with the corresponding sequence
    - $a_{sequence}$ - the most likely ability of OUTs, with respect to which, the result obtained corresponds to a given sequence

For example: p_{011} denotes the proportion of OUTs for *K*=3 where the first test item is not passed, while the second and the third test items are passed. The corresponding ability is a_{011} .

- d_k - denotes the difficulty of k -th test item, $k = 1, ..., K$,
- p_k - denotes the total proportion of OUTs, which successfully passed the k -th test item. The latter is obtained by summing all $p_{sequence}$ for which there is $"1"$ at the k -th sequence' position. For example, for *K*=3, $p_1 = p_{100} + p_{101} + p_{110} + p_{111}$. Note that $\sum_{k=1}^{K} p_k \geq 1$.

2.3. Main assumptions

A1. Two OUT-s, with the same results are considered indistinguishable under given test and have the same - most likely- abilities. This means that the spectrum of possible abilities is limited by 2^K components.

A2. The responses to different test items, applied to the same OUT, are conditionally independent, i.e. response to one test item does not affect any other. In the absence of any prior information about the difficulty of test items, this assumption seems quite plausible.

A3. The response function $P(d|a)$ is self similar (scale invariant). Speaking metaphorically: if the moon (where gravity is six times weaker and therefore the physical ability to jump high is six times greater) would have a breathable (for humans) atmosphere, an athlete could jump six times higher. Mathematically ,this means that

$$P(\lambda d|\lambda a) = P(d|a) \qquad (2)$$

which is possible only if $P(d|a)$ is a function of the ratio d/a between difficulty and ability or *vice versa*. Accordingly, all limitations and solutions are defined and formulated only for relations d/a or a/d . The minimal value $(a/d)_{\min} = 0$. The $(a/d)_{\max}$ is determined in accordance to the test specificity.

A4. There is no chance for OUT with zero ability to overcome any test item, i.e. no pseudo-guessing chance, as in psychometry.

A5. At the initial stage and as the initial approximation, the same abilities are attributed to the same test results. This assumption, of course, can be challenged, but nothing better in the frame of the available information at the initial stage exists. More flexibility is achieved later.

A6. The final distribution of abilities within the tested population must comply with the existing test results sequence distribution under assumed IRF.

3. DISCRETE VERSION OF PROBLEM FORMULATION AND SOLUTION OUTLINE

3.1 Discrete version of the integral equation

Assume that the test consists of K test items of unknown difficulty d_k. The abilities admit 2^K unknown ability values $\{a_{sequence}\}$with probabilities $\{p_{sequence}\}$ that correspond to proportion of OUTs whose test results are consistent with the corresponding sequence. Thus (1) becomes the system of the K equations:

$$p_k = \sum_{by\ \mathbf{all}\ sequences} P(d_k \,|a_{sequence}) \cdot p_{sequence} \qquad (3)$$

for $2^K + k$ unknown abilities and difficulties. For example, for *K*=2 ($2^2 = 4$ sequences) the system (3) becomes

$$p_{00}P(d_1|a_{00}) + p_{10}P(d_1|a_{10}) + p_{01}P(d_1|a_{01}) + p_{11}P(d_1|a_{11}) = p_1$$
$$p_{00}P(d_2|a_{00}) + p_{10}P(d_2|a_{10}) + p_{01}P(d_2|a_{01}) + p_{11}P(d_2|a_{11}) = p_2$$

3.2 Maximal likelihood estimation of $\{a_{sequence}\}$

In order to express $\{a_{sequence}\}$as functions of difficulties, we will use the *maximum likelihood estimation* (MLE) approach, i.e. for every sequence the appropriate likelihood function is built and maximized. The likelihood function is a product of K factors, where the k-th factor (relating to the k-th test item) equals $P(d_k|a_{sequence})$, if in this sequence the k-th tested item was passed successfully and $1 - P(d_k|a_{sequence})$ otherwise. For example, for the forth sequence [0,0,1] of the three-item test in Table 1, the likelihood function *L* takes the following form:

$$L_{001}(a; d_1, d_2, d_3) = (1 - P(d_1|a)) \cdot (1 - P(d_2|a)) \cdot P(d_3|a)$$

Following this approach $\{a_{sequence}\}$as the functions of difficulties are substituted into (3) and this system of equations is solved. As a result we receive *K* difficulty levels, which, after substitution, yield the most likely ability values.

3.3 The final adjustment of ability distribution

The final activity is to find the theoretic distribution of previously found abilities that explains all obtained test results for all sequences.

4. REASONABLE CHOICE OF THE IRF

First, we would like to bring general considerations influencing the choice of a suitable IRF $P(d|a)$. The most primitive approach is to assume that this is the step function that is equal to zero for the ability lower than difficulty and one otherwise. At a more advanced approach IRF is supposed to be some monotonically increasing function of deviation of ability a from the difficulty d, i.e the function of the difference $(a-d)$ [2,3]. Finally, following assumption A.3 and A.4 it must be a function of relative deviation only, i.e. the function of $(a-d)/d$ or $(a-d)/a$ vanishing when $a=0$. Following more or less the classical item response theory (IRT) with its preference given to logistic model and desire not to overload the model by additional parameters we decided to try the following simplest function satisfying the above conditions:

$$P(d|a) = \begin{cases} \dfrac{1+\tanh(1-d/a)}{1+\tanh(1)}, & a \neq 0 \\ 0, & a = 0 \end{cases} \qquad (4)$$

The graphs of (4) as a function of *a* given *d* = 1 and as *a* function of *d* given *a* = 1are depicted in Fig. 1a and Fig. 1b, respectively.

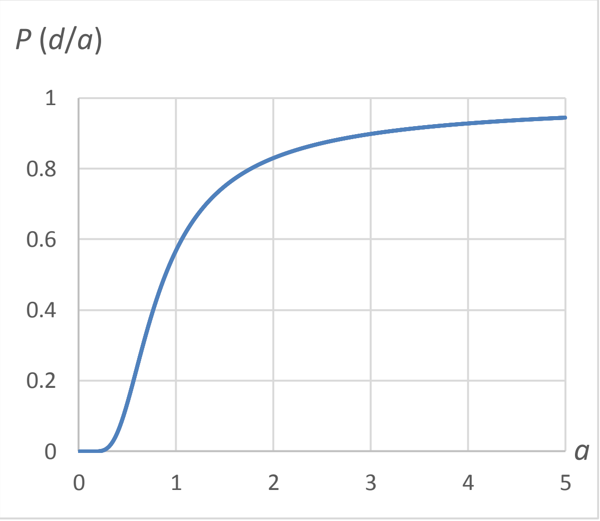

Fig. 1a $P(d|a)$) as a function of *a* given *d* = 1

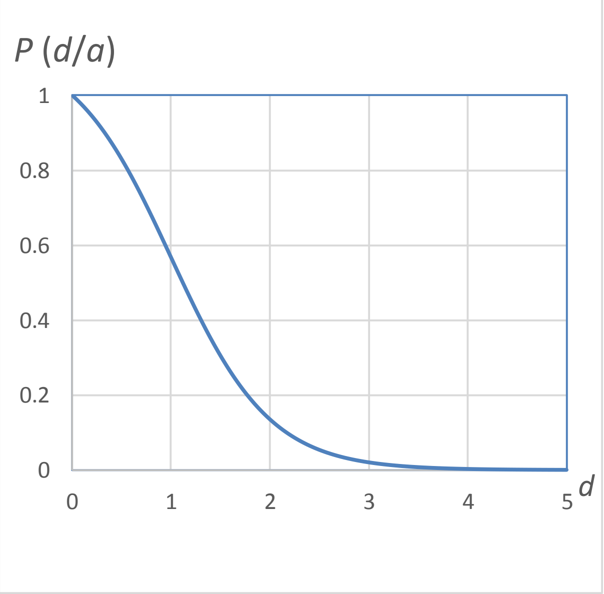

Fig. 1b $P(d|a)$) as a function of *d* given *a* = 1

5. NUMERICAL EXAMPLE

A large batch of detectors passed through three tests: over-rated (cycle acceleration) and two over-stressed: 2) over lighted and 3) under lighted environment. In this example K = 3. The proportions $p_{sequence}$ are given in Table 2.

Table 2 OUTs' test proportions $p_{sequence}$

sequence	$p_{sequence}$
000	0.17
001	0.03
010	0.15
011	0.05
100	0.40
101	0.10
110	0.07
111	0.03

Applying a convergent iterative algorithm based on the described above procedures, the resulting difficulties, normalized by d_1, are : $d_1{*}=1,\ d_2{*}=2.068,\ d_3{*}=2.577$

The resulting abilities, normalized by d_1, are presented in Table 3., showing that the ability increases monotonically with respect to the test difficulty.

Table 3 Resulting abilities

sequence	$a^*_{sequence}$
000	0.00
001	1.06
010	1.60
011	1.61
100	2.12
101	2.40
110	2.99
111	5.00 (upper limit)

The adjusted distribution of these abilities is presented in Table 4. bellow:

Table 4 Resulting distribution of abilities

$a^*_{sequence}$	Sequence relative frequency
0.00	0.06
1.06	0.38
1.60	0.29
1.61	0.22
2.12	0.07
2.40	0.03
2.99	negligible
5.00 (upper limit)	negligible

6. CONCLUSIONS

We propose a new approach to evaluation/measurement of test results, when the test item response is binary and the difficulties of K test items are not known apriori. The evaluation is carried out after the test is performed among some population of objects under test (OUTs), i.e. posteriori, and involves determination of following values:

- difficulty of the test item;
- abilities, assigned to 2^K possible test results;
- anticipated distribution of the above abilities among the tested population

Initially, the most likely ability, implicitly expressed by all unknown difficulties, is attributed/assigned to every possible response sequence of the test, implying equal abilities for the same responses. We assumed that the same response model is valid and applied to all tested objects. In the second stage difficulties themselves are chosen to satisfy the convolution equation (3) relating to the observed distribution of positive responses by different levels of test item difficulty. It is extremely important to emphasize that in this, second step the authors use a scale invariant model in which the probability of a positive response depends only on the ratio between the ability and the difficulty and not on each of them separately. In other words, in the absence of clearly defined units, only the relationships between skills (abilities) and challenges (difficulties) are meaningful and make sense. In the previous step, an assumed ability is uniquely assigned to the test response. In the last, final stage, we relax this relationship, without changing the previously found spectra of difficulties and abilities, with the aim to adjust the ability distribution to observed frequencies. As a result of all the effort, sufficient and consistent solution of the problem, given a determined item response model, is obtained. The authors are aware of all possible refinements and generalizations of the proposed approach, associated with rejection of the restrictions made in the article. Nevertheless, the purpose of this article was to demonstrate principally the feasibility of the proposed approach in solving a very general testing problem. We were guided by the well-known G. Box proverb: "*all models are wrong, but some are useful*".

ACKNOWLEDGMENTS

The authors are grateful to Prof. D. Steinberg for his interest and fruitful discussion of the presented ideas.

REFERENCES

[1] G.J. Engelhard, *Invariant Measurement Using Rasch Model in the Social, Behavioural, and Health Science*. Routledge, 2013.

[2] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment tests*. Paedagogike Institut. Copenhagen, 1960.

[3] B.K. Hambleton, H. Swaminathan, H.J. Rogers, "Fundamentals of Item Response Theory", *Measurement Methods for the Social Sciences,* vol.2, SAGE Publications, Newbury Park, London, New Delhi, 1991.