Measurement 78 (2016) 348–357

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Testing and evaluating one-dimensional latent ability

Vladimir Turetsky, Emil Bashkansky *

Ort Braude College, Karmiel, Israel

ARTICLE INFO

Article history:
Available online 11 June 2015

Keywords:
Testing
Latent ability
Difficulty
Item response
Maximal likelihood

ABSTRACT

A new approach to evaluation of binary test results when checking a one-dimensional ability is proposed. We consider the case where a qualitatively homogeneous population of objects is tested by a set of non-destructive test items having different, but unknown beforehand levels of difficulty, and we need to evaluate/compare both the intrinsic abilities of these objects and the level of difficulty of the test items. We assume that the responses to different test items, applied to the same part, do not affect one another and the same scale invariant item response model applies to all members of the tested population of objects under test (OUTs). OUT can mean an electronic component, examinee, program unit or material under test, etc. An algorithm for solving the problem, applicable for engineering testing, is proposed. It combines item response theory, maximum likelihood estimation, method of flow redistribution and other methods. Numerical example is presented.

1. Introduction

The English language contains hundreds of words directly or indirectly describing the different types of overt and hidden *abilities*: from cognitive as, for example, memory and attention, to purely technical, such as reliability, stability, capability, availability, durability, portability, and reusability. In this paper, we deal only with the basic and simplest issue of so-called one-dimensional ability, when the *test item* performance of the object under test (henceforth abbreviated OUT) can be explained by a single latent ability. 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/-compare both the intrinsic abilities of these OUTs and the difficulties of the test items. This type of set hereinafter will be called the test and it can include any – but must be the same for all OUTs – number of test items. For instance, in the psychometrics test the OUT is an examinee, a separate question on the exam is a test item and the examination as a whole is a test. Usually, it is assumed [1] 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). It is also assumed that the inherent ability of the OUT is independent of the test item difficulty. Homogeneity here means that the same item response model is applied to all members of the population, but in any case does not imply equality of the tested abilities among these members.

Even in such a simplified model the matter of correct and effective evaluation of test results has not been resolved completely and is still a subject of discussion in psychometrics and educational measurement [2]. An extensive study of latent ability modeling and evaluation in education was pioneered by [1], who proposed a well-known model of an interconnection between the test item difficulty, the examinee's ability and the test result based on a standard logistic distribution. The Rasch model was extensively studied and extended during the last decades (see, e.g., [3–7] and references therein). The problem of estimating the Rasch model

* Corresponding author.
E-mail address: ebashkan@braude.ac.il (E. Bashkansky).

http://dx.doi.org/10.1016/j.measurement.2015.05.048
0263-2241/

parameters is tackled mainly by using one version of the maximal likelihood estimation [8].

The problem, however, is discussed to a much lesser extent in engineering, which prefers to deal with quantifiable test results, estimated on a predetermined scale of difficulties (e.g., life time testing). This state of affairs seems a little strange, since in the broader context, testing can be subjected to any property and any OUT: people, program units, electronic components, materials, network connectivity, etc. Moreover, engineering objects of interest are more predictable and less variable, being free from purely human restrictions. The technical test population can often easily be established as more or less uniform, for instance when all the parts belong to the same production batch. In view of this, it seems desirable to develop some unifying/standardized approach for evaluating test results of such objects, when the difficulty of test items is unknown beforehand. The combination of two tests items – over-stressed and overrated [9] – can serve an example of such a test as well as testing including a wider variety of test items.

We propose an algorithm for test result evaluation applicable to a broad spectrum of engineering tests satisfying the model assumptions described below in Section 2. 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 OUTs according to some cumulative distribution function (cdf) $F(a)$. 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 (or level of difficulty) of the test item in relation to the studied ability. For the purpose of illustration only, let us, for instance, consider an athlete's physical fitness as a and the height of the bar as d. We assume that there is some known or supposed function $p(d|a)$, customarily called the *item response function* (IRF), which expresses the probability that the OUT with ability a will successfully overcome the test item of difficulty d [8]. It is natural to assume that the greater d is, the less is this probability for every given a, i.e.,

$$\frac{\partial p(d|a)}{\partial d} < 0. \tag{1}$$

Then, the proportion $p(d)$ of OUTs that successfully overcome the test item having difficulty d is

$$p(d) = \int p(d|a)dF(a). \tag{2}$$

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 the test item is, the less OUTs pass it successfully.

It would seem that in the mathematical sense (2) is a Fredholm integral equation of the first kind, representing a classic measurement inverse problem [10,11], in which one wants to restore the measured value 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' level of difficulty in our case is unknown beforehand, and neither is $p(d)$. 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 difficulty $d_k, k = 1, \ldots, K$. Every OUT is tested independently, i.e., results of one OUT do not affect the results of another. *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 $d_{k+1} \geqslant d_k$. In total, there exist 2^K possible test results and this resolution does not allow us to suggest more than 2^K distinct levels of ability. It is obvious that the greater is K, the better is the resolution. Using "0" to indicate failure and "1" to indicate the successful completion of a 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 of $K = 3$ items, all possible results are presented by their respective *binary codes* as shown below in Table 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 the 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 kth test item, $k = 1, \ldots, K$;
- p_k – denotes the total proportion of OUTs, which successfully passed the kth test item. The latter is obtained by summing all p_{sequence} for which there is "1" at the kth sequence' position.
  For example, for $K = 3, p_1 = p_{100} + p_{101} + p_{110} + p_{111}$. Note that $\sum_{k=1}^{K} p_k \geqslant 1$.

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