Measurement 122 (2018) 20–26

Contents lists available at ScienceDirect

Measurement

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

Binary test design problem

Vladimir Turetsky[a], David M. Steinberg[b], Emil Bashkansky[c,*]

[a] ORT Braude College, Department of Applied Mathematics, P.O. Box 78, 51 Snunit, Karmiel 2161002, Israel
[b] Department of Statistics and Operations Research, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[c] ORT Braude College, Department of Industrial Engineering & Management, P.O. Box 78, 51 Snunit, Karmiel 2161002, Israel

ARTICLE INFO

Keywords:
Binary test
Ability
Difficulty
Optimization criteria

ABSTRACT

Unlike in traditional measurement methods, in binary testing each test item provides only one bit of information. In view of limited test resources, effective planning of the test is crucial. In this article, the general problem is formulated from the metrological point of view for a high variety of objects under test and a homogeneous item response function. Different optimization criteria are reviewed for one-item testing (single and replicated), and their advantages and disadvantages are discussed. The article concludes with preliminary recommendations for how to plan a binary test.

1. Introduction

Binary tests have many practical purposes and in recent years, substantial progress in the analysis and interpretation of binary test results in engineering applications has been achieved [1–12]. Binary tests are not very informative, since they supply very few bits of information, and therefore the issue of planning/design of these tests becomes especially important. In this paper, we restrict ourselves to the simplest issue of measuring a unidimensional ability a, when the *test item* performance of the *object under test* (OUT) can be explained by a single latent ability. OUT can mean an electronic or mechanical component, process, data, program unit or material under test, etc. A statistical hypothesis test can also serve as an example of a binary test [5]. A special case is created when the OUT is a person, whose cognitive/physical abilities are studied by psychometrics ([16,17] and references therein), kinesiology (e.g., [18]) and other disciplines. We do not consider ourselves experts in such a complex field, but hope that some ideas, presented below, may be useful there also. In recent years, one can observe a promising mutual diffusion of ideas between metrological, engineering and psychometric test approaches [19–24].

The test consists of a set of K non-destructive test items. Every test item response is scored on a binary scale (pass/fail) and the target is to evaluate the intrinsic ability of the tested OUT. Usually, it is assumed that 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). Given a specific *item response function* (IRF) model, i.e., the probability that the OUT with ability a successfully overcomes the test item having difficulty d, assessment of the tested ability is usually based on the principle of maximum likelihood estimation (MLE). When the levels of test item difficulties $(d_1,d_2,...,d_K)$ are known beforehand, the solution of the problem is relatively easy, but when the number of OUTs is bounded and the levels of difficulties are unknown beforehand, the resulting analysis involves significant computational difficulties. Nevertheless, in principle, the problem of measuring/scoring the tested abilities is solvable.

However, when we think how to allocate test resources optimally, we are faced with several problems: how to choose the levels of test item difficulties, how many repetitions to perform for every level, what is the criterion of optimality, etc.

Perhaps, the most serious attention to these questions was given in the field somewhat remote from engineering: in psychometrics and educational measurement [25–34], when developing banks of test items and sequential computer adaptive testing (CAT). In light of our article, approaches using various information aspects for planning a test and estimating the uncertainty of its result are of main interest [16,23 and references therein]. Nevertheless, the overwhelming majority of the abovementioned studies are based on properties of the Rasch item response model [25] and its latest modifications [30,33], recognized as the main model in psychometrics.

The difference between psychometrical, technical, financial, statistical, physical and other tests consists in the models used to describe the specific item response function (IRF). In technical, financial, statistical, physical testing, the response models may differ significantly and no longer have the remarkable properties of the Rasch model [35], while acquiring some other properties (such as self-similarity, for example [5,6,12]) The mathematical expressions of the IRF in most cases are quite distinct. Therefore, it makes sense to discuss the problem of binary test planning from the most general principles point of view.

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

https://doi.org/10.1016/j.measurement.2018.02.031
Received 22 September 2017; Received in revised form 16 February 2018; Accepted 17 February 2018
Available online 26 February 2018

Table 1
Ability vs. difficulty.

Ability	Difficulty
Of a detector to detect a defect	The size of the defect
Explosiveness	Trip wire current
Mechanical strength	Applied load
Of a control chart to detect a shift in the mean of the process	Shift value
Of an athlete to physically jump upwards	Bar height

This paper offers possible approaches to the test planning problem not limited to the specific item response model, although the illustrative examples are tied to a self- similar model, described in the following section.

2. Item response function (IRF) model

The two main components of binary testing are:

(1) The object under test (OUT)
(2) The test item (TI).

An OUT is characterized by its *ability* (to be precise: level of ability), a. This can be the ability of the detector to detect a defect, the explosive sensitivity of a material, mechanical strength, the ability of a control chart to detect a shift in the mean of a process, an athlete's physical ability to jump upwards, etc.

TI is characterized by its *difficulty* (to be precise: level of difficulty), d. This can be, accordingly, the size of the defect, the current in a trip wire, applied load, shift value, bar height, etc. (see Table 1).

Except in the cases when the result of testing is predetermined by the values of ability and difficulty, the relationship between them is described by the so-called *item response function* (IRF) – $P(d,a)$. This expression represents the probability that the OUT with ability a will successfully pass the test item of difficulty d. Each specific area of study has its own, and often not only one, IRF model. Starting in psychometrics, the IRF concept has spread to a variety of areas including social and behavioral sciences [22,23], educational measurement [10,11,14], medicine [17], quality engineering [6–8] and other practical testing problems [36].

In most cases, ability and difficulty (directly or through some transformation) can be brought to the same measuring scale. Under such conditions, so-called scale invariant models [12] are usually used. In these models, changing the unit of measure does not affect the probability of successfully passing the test item. This means that the IRF satisfies the Euler functional equation $P(\lambda d,\lambda a) = P(d,a)$ for any $\lambda > 0$ implying that the IRF is a homogeneous function of the zeroth order. It is known [15] that the only non-trivial solution of this equation is of the form $P(d/a)$, i.e., the IRF is a function of the ratio between difficulty and ability and not of each of them separately. For certainty, our further reasoning is based on the following IRF of this type, which originated in statistical process control [5] (see Appendix for its derivation):

$$P(d/a) = 1-\Phi\left(\sqrt{2}\cdot\ln\frac{d}{a}\right), \tag{1}$$

where Φ denotes the standard normal cumulative distribution function. According to this model, $0 \leqslant a,\ d \leqslant 1$, where $d = 1$ indicates a *placebo*. Common with the Rasch IRF is that $P(d/a) = 0.5$ when $d = a$. In Fig. 1, this IRF is depicted as a function of the ratios a/d and d/a.

3. Use of preliminary (prior) information

What happens when we have information prior to testing and when we do not?

- When there is a complete lack of any information regarding the measured ability—the ability is evaluated based on test results only.
- When we have some prior information, some assumptions regarding the tested ability, expressed by the *prior* probability density function $f(a)$, can be made. For example, considerations based on historical data can be involved. In this case, the Bayesian approach of re-evaluating the prior hypothesis is required. The exception is the situation in which the information received from the test is much greater than what was assumed about the ability before testing. In this case, preliminary information can be neglected.

4. Notation prelude

Suppose that K test items having the same or different difficulty levels $(d_1,d_2,d_3,\ldots,d_k,\ldots d_K) = \vec{\boldsymbol{d}}$ are applied to the OUT. The result of the whole test can be presented as a data vector/sequence of ones and zeros, such as $\vec{\boldsymbol{s}} = (s_1,s_2,s_3,\ldots,s_k,\ldots s_K) = (1,0,1,\ldots,0,\ldots 0,\ldots 1)$, where 1 means that the corresponding test item passed successfully and 0 otherwise. Clearly, there are 2^K possible test results and, respectively, 2^K possible

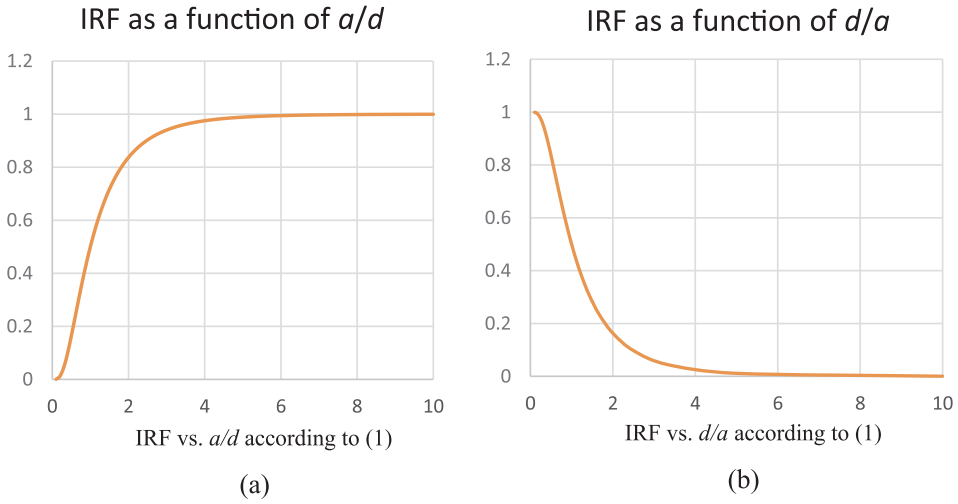

Fig. 1. (a) IRF vs. a/d according to (1). (b) IRF vs. d/a according to (1).