Ability evaluation by binary tests: Problems, challenges & recent advances

E Bashkansky\textsuperscript{1} and V Turetsky\textsuperscript{2}
ORT Braude College of Engineering, P.O. Box 78, Karmiel, 2161002, Israel
\textsuperscript{1}Industrial Engineering and Management Department
\textsuperscript{2}Department of Applied Mathematics

E-mail: ebashkan@braude.ac.il

Abstract. Binary tests designed to measure abilities of objects under test (OUTs) are widely used in different fields of measurement theory and practice. The number of test items in such tests is usually very limited. The response to each test item provides only one bit of information per OUT. The problem of correct ability assessment is even more complicated, when the levels of difficulty of the test items are unknown beforehand. This fact makes the search for effective ways of planning and processing the results of such tests highly relevant. In recent years, there has been some progress in this direction, generated by both the development of computational tools and the emergence of new ideas. The latter are associated with the use of so-called “scale invariant item response models”. Together with maximum likelihood estimation (MLE) approach, they helped to solve some problems of engineering and proficiency testing. However, several issues related to the assessment of uncertainties, replications scheduling, the use of placebo, as well as evaluation of multidimensional abilities still present a challenge for researchers. The authors attempt to outline the ways to solve the above problems.

Introduction

Qualitative testing aimed to measure the latent ability and resulting in a binary output of type “good/bad”, “pass/fail”, “true/false”, “yes/no”, “positive/negative”, “detected /non detected”, “1/0” etc. is often used in various fields of science, educational measurement, quality control, industry and healthcare, including analytical chemistry and microbiology.

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 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, (known or unknown beforehand) levels of difficulty, and we need to evaluate/compare the intrinsic abilities of these OUTs and the levels of difficulty of the test items if the latter are unknown beforehand. 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. We assume [1] that the test item response is estimated on the binary scale base (“good/bad”, “pass/fail”, “true/false”, “yes/no”, “positive/negative”, “detected /non detected”, “1/0” etc.) 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. Such qualitative testing is often used in various fields of science, educational measurement, quality control, industry and healthcare, including analytical chemistry and microbiology [2-10]. As the response to each test item provides only one bit of information per OUT, for a full assessment of an OUT’s ability (hereafter denoted by \(a\)), the test usually comprises a series of test items presenting various levels of difficulty (hereafter denoted by \(d\)). These levels can be known or unknown beforehand. The latter situation presents a case that is more complicated to analyze. The search for effective ways of planning and processing the results of such tests is highly relevant. In recent years, there has been some progress, generated by both the development of computational tools and the emergence of new ideas, which will be presented in this article.

2. Main definitions and assumptions

2.1 Definitions
Consider \(m\) OUTs taking part in a testing procedure. Every OUT has to pass the test consisting of a set of \(n\) test items, having different levels of difficulty. Let \(a\) denote the ability of OUT \(i\) \((i=1,2,...I)\) and \(d\) the difficulty of the test item \(j\) \((j=1,2,...J)\) in relation to the studied ability. In some experiments every OUT can be tested by the same test item repeatedly. For each intersection \(i \times j\) we will use \(r_{ij}\) to denote the number of such repetitions, \(r_{ij}^{+}\) for the amount of positive results, \(r_{ij}^{-}\) for the amount of negative results and \(f_{ij} = \frac{r_{ij}^{+}}{r_{ij}}\) for the relative frequencies of positive results obtained in these repetitions. Also, we use 1 to indicate a positive result and 0 otherwise. Accordingly, the results of the testing can be presented by an \(I \times J\) matrix \(R\), whose elements \(r_{ij}\) may change from zero to \(r_{ij}^{+}\).

Thus, results of testing every OUT by three test items can be presented as in Table 1.

| OUT1 | Test item 1 | Test item 1 | Test item 2 | Test item 2 | Test item 3 | Test item 3 |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| \(r_{11}^{+}\) | \(r_{11}^{-}\) | \(r_{12}^{+}\) | \(r_{12}^{-}\) | \(r_{13}^{+}\) | \(r_{13}^{-}\) |

Further, let \(P(d | a)\), customarily called the item response function (IRF), express the probability that the OUT with ability \(a\) will successfully pass the test item having difficulty \(d\). The following reasonable and quite plausible main assumptions concerning the testing can be made [1].

2.2 Assumptions concerning the test results

2.2.1 The responses to different test items, applied to the same OUT, are conditionally independent, i.e., the response to one test item does not affect any other. This is admissible, if no direct link between the difficulty of the test item and the inherent ability of the OUT exists. In other words, the difficulty of the test item does not affect the inherent ability of the OUT.

2.1.2 Testing is organized in such a way that, thanks to randomization, OUT does not know whether it has to pass the same or a different test item, so the responses to repeated test items, applied to the same OUT, are also conditionally independent as in the Bernoulli scheme with probability of “success” \(P(d | a)\).

2.2.3 Two OUTs with the same results are considered indistinguishable under the given test and judged to have the same—most likely—ability.
2.3 Scale invariant item response model
We assume that ability and difficulty (directly, or through some transformation) may be brought to the same measuring scale. Then, applying considerations of dimensional analysis, we conclude that the dimensionless response function should be scale invariant. Mathematically, this means that for any reasonable \( \lambda > 0 \):

\[
P(\lambda d \mid \lambda a) = P(d \mid a),
\]

which is possible only if \( P(d \mid a) \) 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 \( a/d \).

2.4 Novel process control based IRF
Let us interpret a standard statistical process control procedure [11] as the ability of the \( \bar{x} \) control chart to detect a shift in process quality (a shift from the in-control value), when the standard deviation is assumed known and constant. Denote the magnitude of the shift, measured in units \( \sigma / \sqrt{n} \) (where \( n \) means the sample size), as the \( k \) and a distance between control limit and centre line, measured in the same units, as \( L \). Usually, the \( \bar{x} \) control chart has two limits, thus essentially, it combines two detectors. To clarify, we will take into account only one of the limits—the upper one—for example. Then, according to [11], the probability of detecting the shift up of magnitude \( k \) is given by \( 1 - \Phi(L - k) \), where \( \Phi \) denotes the standard normal cumulative distribution function. Now, define:

- **Ability** \( a \) of the one-sided control chart to detect the change: \( a = e^{-k} \) (the farther from the center the control limit is, the lower the sensitivity of the control chart is),

- **Difficulty** \( d \) of detecting the shift: \( d = e^{-k} \) (the larger the shift is, the easier is to detect it and vice versa).

Both \( a \) and \( d \) are defined in the range of zero to one; nevertheless, their ratio can take any value. The maximal level of difficulty \( d = 1 \) is achieved for the case when the shift essentially does not exist \( (k = 0) \) so a positive response signifies a false alarm. Hereafter, such test items will be called placebos.

Thus:

\[
P(d \mid a) = 1 - \Phi (-\sqrt{2} \cdot \ln a + \sqrt{2} \cdot \ln d) = 1 - \Phi (\sqrt{2} \cdot \ln \frac{d}{a}).
\]

In our previous publications [1], we used a Rasch similar IRT:

\[
P(d \mid a) = [1 + \tanh(1 - \frac{d}{a})] / [1 + \tanh 1].
\]

Figures 1.a and 1.b demonstrate that the IRF presented in Formulas (2) and (3) are very close one to another. Hereafter, we will mainly use the IRF in the form provided by Formula (2).
3. Review of the types of problems and ways to solve them

We now discuss what types of problems researchers may face and the possible approaches for solving them.

3.1 Levels of difficulty of the test items are known beforehand

Suppose the levels of difficulty of the test items \( d_1, d_2, \ldots, d_J \) are given and results of testing every \( i \)-th OUT are represented in a form similar to Table 1. Then the problem of estimating \( i \)-th’s ability can be resolved by applying the maximum likelihood estimation (MLE), i.e., by maximizing the likelihood function:

\[
L(a_i; d_1, d_2, \ldots, d_J) = \prod_{j=1}^{J} P(d_j | a_i)^{f_{ij}} \cdot [1 - P(d_j | a_i)]^{1 - f_{ij}}. \tag{4}
\]

It seems simple, but the problem is that if the test results themselves are not reliable, improbable and self-contradictory—the probability of passing the test items predicted by \( P(d_j | a_i) \) may be very far from the observed relative frequencies \( f_{ij} = r_{ij}^{(+)} / r_{ij} \). In this case, a chi-square or similar test can serve as indicator of confidence in the estimated ability.

In example presented below in Table 2, \( d_1 = 0.2 \); \( d_2 = 0.6 \); \( d_3 = 1.0 \) (placebo) and the most likely value of the ability equals 0.4. Predicted and observed frequencies are very close and \( \chi^2 = 0.008 \) is very small.

If for the same test items the obtained results appear as presented in Table 3 (only results of placebo test item are interchanged), the most likely value of the ability equals 0.53. Predicted and observed frequencies are very far and \( \chi^2 = 39.3 \) is very large. This result is not because of a bad assessment. It is the consequence of the inconsistency/dissonance of the testing results themselves: indeed, 9 from 10 positive responses to placebo looks beyond common sense!
Table 2. Results of testing OUT $i$ given $d_1 = 0.2$; $d_2 = 0.6$; $d_3 = 1.0$ (placebo)

| OUT $i$ | Test item 1 $r_{i1}^{(+)} = 21$ | Test item 1 $r_{i1}^{(-)} = 4$ | Test item 2 $r_{i2}^{(+)} = 10$ | Test item 2 $r_{i2}^{(-)} = 26$ | Test item 3 $r_{i3}^{(+)} = 1$ | Test item 3 $r_{i3}^{(-)} = 9$ |
|---------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Pred.   | 20.91                         | 4.09                          | 10.19                         | 25.81                         | 0.98                          | 9.02                          |

Table 3. Observed vs. expected/predicted frequencies

| OUT $i$ | Test item 1 $r_{i1}^{(+)} = 21$ | Test item 1 $r_{i1}^{(-)} = 4$ | Test item 2 $r_{i2}^{(+)} = 10$ | Test item 2 $r_{i2}^{(-)} = 26$ | Test item 3 $r_{i3}^{(+)} = 9$ | Test item 3 $r_{i3}^{(-)} = 1$ |
|---------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Pred.   | 22.90                         | 2.10                          | 15.49                         | 20.51                         | 1.85                          | 8.15                          |

3.2 Levels of difficulty of the test items are unknown beforehand and OUT population is very large

Suppose we consider a very large population of OUTs tested without repetitions (i.e., $r_{ij} = 1$) by $J$ test items with unknown beforehand levels of difficulty. “Large” means that it is big enough to consider a proportion of the OUTs that have successfully passed some test item as a probability of its successful passing by being randomly selected from the OUT population. So, if the studied ability $a$ is distributed among the tested population of OUTs according to some probability density function $f(a)$, then the proportion $f(d)$ of OUTs that successfully overcome the test item having difficulty $d$ is expected to be some convolution of $f(a)$ with $P(d|a)$:

$$p(d) = \int P(d|a)f(a)da.$$  \hspace{1cm} (5)

Here, the OUTs with the same test results are considered indistinguishable under a given test and judged to have the same most likely abilities. Proposed in [1], the iterative procedure allows us, in this case, to find both spectra: $J$ difficulties and $2^J$ abilities, as well as the distribution of these abilities among the studied population.

3.3 Levels of difficulty of the test items are unknown beforehand and the OUT population is limited

For this case, the OUTs with the same test results are also considered indistinguishable under a given test and judged to have the same most likely abilities. The likelihood function takes the form:

$$L(a_1,a_2,\ldots,a_I; d_1,d_2,\ldots,d_J) = \prod_{i=1}^{I} \prod_{j=1}^{J} P(d_j|a_i) r_{ij}^{(+)} \cdot \left[1 - P(d_j|a_i)\right] r_{ij}^{(-)},$$  \hspace{1cm} (6)

and depends on $I+J$ variables. Although it is usually much less than in $I + 2^J$, as in 3.2, the problem of finding the maximum in $I+J$ space is rather serious, a bit easier because apparently this maximum is unique. For example, let us suppose that the results of the test are those that are shown in Table 2, but except for $d_3 = 1$ (placebo) the rest of the difficult levels $(d_2, d_4)$ are unknown. By maximizing (6), we get the following: $d_1 = 0.2, d_2 = 0.613$; $a_i = 0.4041$. For more than one OUT, the problem becomes more complicated and encounters computational difficulties [12].
4. Challenges

The problem of binary test evaluation in the case when the levels of difficulty of the test items are unknown beforehand and the amount of tested OUTs is limited and large remains a challenge for researchers dealing with binary testing in different fields of science and engineering. A problem of optimal allocation of the experimental resources is also yet unsolved.

References

[1] Turetsky V and Bashkansky E 2016 Testing and evaluating one-dimensional latent ability Measurement 78 348–57.
[2] AOAC (Association of Official Agricultural Chemists) Official Methods of Analysis 2013 Guidelines for Validation of Qualitative Binary Chemistry Methods, International Stakeholder Panel on Alternative Methods (ISPAM).
[3] Bashkansky E and Gadrich T 2013 Some statistical aspects of binary measuring systems. Measurement 46 1922–27.
[4] Wehling P, LaBudde RA, Brunelle SL, Nelson MT 2011 Probability of detection (POD) as a statistical model for the validation of qualitative methods. J AOAC Int. 94(1) 335–47.
[5] Wilrich C and Wilrich P 2009 Estimation of the POD function and the LOD of a qualitative microbiological measurement method J AOAC Int. 92(6) 1763–72.
[6] Wilrich P 2010 The determination of precision of qualitative measurement methods by inter-laboratory experiments Accreditation and Quality Assurance 15 439–44.
[7] Altman G, Bland M 1994 Diagnostic tests. 1: Sensitivity and specificity. BMJ 308 (6943) 1552.
[8] Uhlig S, Krugener S, Gowik P 2013 A new profile likelihood confidence interval for the mean probability of detection in collaborative studies of binary test methods Accreditation and Quality Assurance 18 367–72.
[9] Pendrill L and Fisher W 2015 Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of numbers Measurement 71 46–5.
[10] Uhlig S et al. 2015 Qualitative PT data analysis with easy-to-interpret scores Accreditation and Quality Assurance, 20 347–353
[11] Montgomery D 2009 Introduction to Statistical Quality Control 6th ed. (New York: John Wiley & Sons Inc.) 247 eq. (6.19).
[12] Bashkansky E and Turetsky V 2016 Proficiency testing: binary data analysis Accreditation and Quality Assurance, published online 21 April 2016, DOI: 10.1007/s00769-016-1208-x