How to create an Adaptive Test

Writing an exam is not a straight-forward task. It involves more than just cramming as many challenging questions as possible. The point of having an exam, after all, is to evaluate and discriminate against a population of test-takers. You need a way to know who is performing well in a given criterion and later, group them for psychometric needs.

There's a slight problem to this end—making a fair exam is hard. It is pretty difficult to create an evaluation for a large number of people. You need to ensure that the scores, however they might become, will reflect a person's objective capability.

For a test to be meaningful, it must possess two qualities: Reliability and Validity. Reliability is consistency across time, across items, and across researchers. Validity is the extent to which the scores actually represent the variable they are intended to.[1]

CTT

To build a test like this, psychometricians usually rely on well-established theoretical frameworks. In the past, we relied on Classical Test Theory, or CTT for short. It postulates that every score a person obtains on a test, the Observed Score (X), is a composite of two components: their True Score (T) and a component of random Error (e).

$$X = T + e$$

The True Score (T) is a hypothetical concept. It represents the average score an individual would achieve if they were to take the test an infinite number of times, with the random errors cancelling each other out. We can never directly observe the true score; we can only estimate it from the observed score.

The Error (e) component represents all of the random, unsystematic factors that cause an observed score to deviate from the true score. These factors can include a test-taker's mood, fatigue, luck in guessing, or minor variations in the testing environment.

Despite its influence and intuitive appeal, CTT has significant limitations. Its most critical shortcoming is that its statistics are sample-dependent. An item's

[1] Price, P. C., Jhangiani, R., & Chiang, I-Chant A. (2015). *Research Methods in Psychology* (2nd Canadian ed.), Chapter 5: Reliability and Validity of Measurement. BCcampus Open Education.

difficulty (the proportion of people who get it right) and a test's overall reliability are intrinsically tied to the specific group of people who took that particular test. This makes it difficult to compare scores from different test forms or to compare the ability of students who took different tests.

Furthermore, CTT treats the test as a single, monolithic entity, focusing almost exclusively on the total score. It provides little information about how individuals perform on specific items, and it cannot easily separate the characteristics of the test-taker from the characteristics of the test itself. These limitations paved the way for a more sophisticated model—Item Response Theory.

IRT

The item response theory (IRT) $P_{ij}\left(\theta_j,\ b_i,\ a_i,\ c_i\right)$ models the relationship between a person's underlying ability, other traits, and their probability of answering a single item correctly. This enables researchers to better understand the behaviour of the test-taker and the interaction between the test, which will be important later on to create an adaptive test model.

Now, to visualise the IRT model, we use something called the Item Characteristic Curve (ICC, for short). This sigmoid-shaped curve is derived from the model and represents the probability of choosing a correct item with respect to the ability.

There are three common types of IRT models:

$$P_{ij}\left(\theta_j, b_i\right) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

The 1-PL (or Parameter Logistic) model is the simplest, with only a Difficulty parameter. This parameter shifts the ICC to the right for more difficult questions, and to the left for easier questions. You can see the majority of people who are more likely to answer correctly shifts to the positive side of the ability scale as the question becomes more difficult.

$$P_{ij}\left(\theta_j, b_i, a_i\right) = \frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]}$$

The 2-PL adds another Item Discrimination parameter. In the ICC curve, it is demonstrated by the steepness. Here, a small change of ability can result in a greater positive increase in probability. In this case, the question is effectively separating test-takers over a very small change in their ability.

$$P(\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]}$$

The 3-PL adds another Item Guessing parameter. It is demonstrated by the asymptote of the ICC curve as ability approaches negative infinity. This restricts the probability of endorsing a correct response when the ability of the respondent approaches negative infinity.

Difficulty calibration

Okay, it might be a bit disconnected at first to see those formulae, but let's show how it helps us in measuring test-takers' ability. For the sake of demonstration, we only use the 1-PL IRT model (empirically, it is powerful enough to estimate a test-taker's ability with a smaller sample size). This is the sigmoid function of the ability subtracted by difficulty, which converts the subtraction from negative infinity to positive infinity to the range of 0 to 1.

$$p\left(Y_{i,j}|b_i\right) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} = \sigma(\theta_j - b_i)$$

I have changed the symbols a bit, but they mean the same thing, with Y as the response matrix of either True or False. If we have the probability of answering a question correctly, and the question's difficulty, we can approximate the test-taker's ability by substitution.

There are many ways to solve this equation statistically for ability, but today I will use Machine Learning to estimate both the difficulty and ability. What you will need is a fixed number of random test-takers. To make sure that it is representative of the general population, you have to make sure that they are randomly selected. We will also need a question bank and encode it with the corresponding True / False from collected responses. We are performing this experiment on dichotomic tests, or those that only have the answers as either True or False.

Our goal is to use the unknown innate ability from test-takers and their probability of answering a given question correctly to estimate the difficulty. But how do you use something that is unknown? Well, we know its assumed distribution, that is, standard normal. It is reasonable to assume that the test-takers' abilities, albeit unknown, follow a standard normal distribution. The majority of the population is mediocre, while only a small fraction of test-takers exhibit the extremely high or low abilities found in the tails of the distribution. We can just

generate them randomly following such distribution to find out the difficulty and average out by the number of random samples to get the result.

This is called a Monte Carlo approximation, a powerful tool to use randomness to solve a problem. Here, we used the assumed distribution to generate reasonable samples. Now, we just need to put this through a learning algorithm.

The objective function of this process is to maximise the probability of observing the actual data we have. In other words, we want to find the model parameters (like the question's difficulty) that make the real-world outcomes seem as likely as possible. We average them out to complete the Monte Carlo approximation and put the negative sign in front to turn it into a loss function.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log P(Y_{i,j}|p_i)$$

To actually drive the learning process, we need something called an optimisation algorithm. Usually, we will use Gradient Descent, which is a simple and effective way to find the minimum of a function. For this example, we will use L-BFGS, which is a variant that can handle large-scale problems and helps find the minimum faster.

$$\text{Gradient Descent: } \theta_{k+1} = \theta_k - \alpha_k \nabla f(\theta_k)$$

$$\text{L-BFGS: } \theta_{k+1} = \theta_k - \alpha_k [B_k]^{-1} \nabla f(\theta_k)$$

We can now use this loss function to estimate the difficulty of the questions. Finally, by using the fitted difficulty, we can repeat the same process to estimate the ability for all test-takers.

Adaptive Question Selection

Through calibrating the questions' difficulty, along with estimating individuals' abilities, we have laid the groundwork for creating an adaptive test. Now, our goal is to deliver only the questions that will yield the most information about the test-taker's ability.

A question is most informative when its difficulty level is close to the test-taker's ability level. Asking a very easy question to a highly capable individual tells you very little, as does asking an impossibly hard question to a weak individual.

This will ensure that only the most optimal questions are asked, reducing time, cost, experience and increasing the accuracy of the test.

We can employ the Maximum Fisher Information (MFI) strategy to select the most informative question. This strategy selects a question that the model has roughly a 50% chance of answering correctly, as this is where the most uncertainty is resolved.

Revisiting the 1-PL model, the predictive probability that the current test taker answers item j correctly is

$$p_j \equiv p\big(Y_{\text{new},j} = 1 \mid \theta^t_{\text{new}},\ \hat{z}_j\big) = \sigma\big(\theta^t_{\text{new}} - \hat{b}_j\big),$$

where $\hat{z}_j$ denotes the current estimates for the item parameters and $\hat{b}_j$ is the item difficulty.

The Fisher information of item j for a Bernoulli response under the Rasch model is

$$\mathcal{I}\big(\theta^t_{\text{new}};\ \hat{z}_j\big) = p_j\,(1 - p_j). \tag{1}$$

MFI chooses the next item by maximizing this information at the current ability estimate and then removes it from the remaining pool Q^t:

$$\hat{z}_j^{*t},\ q_j^{*t} = \underset{\hat{z}_j:\ q_j \in Q^t}{\arg\max}\ \mathcal{I}\big(\theta^t_{\text{new}};\ \hat{z}_j\big)\,, \qquad Q^{t+1} = Q^t \setminus \{q_j^{*t}\}. \tag{2}$$

After administering the selected item and observing the response, the ability estimate is updated, e.g., by maximum likelihood over all t administered items:

$$\theta^{t+1}_{\text{new}} = \underset{\theta}{\arg\max}\ \sum_{j=1}^{t} \log p\big(Y_{\text{new},j} \mid \theta,\ \hat{z}_j\big). \tag{3}$$

This cycle of (i) pick by (2), (ii) observe, (iii) update by (3) repeats until a stopping rule is met (e.g., target standard error, max items, or budget).[2]

[2] See, e.g., Baker (2001) and van der Linden et al. (2000) for background on item selection and Fisher information in adaptive testing.