ON PUBLIC-QUESTION TESTS

GANG XIAO

This article describes a new test system, based on computer-generated random variation of questions, that has many advantages over the traditional test method.

Introduction

A *test* is a system designed to evaluate the knowledge or competence of a certain number of people (candidates). Questions are asked to the candidates, and scores are attributed to each according to the rate of correct answers.

In a traditional test, the same set of questions must be given to all the candidates, in order to ensure equality of chances among the candidates. This brings about many difficulties for its realization: test sessions must be synchronized for all the candidates and the questions carefully kept secret before the session openings, and copies between neighboring candidates are always very hard to eliminate. While real equality of chances can never be achieved, as a candidate who happens to have seen the selected questions better will be privileged. Also, designing a special question set for each test means that the result is not quantifiable, that is, the score of one test cannot be compared with that of another one, because both depends on the question set whose selection is usually subjective.

On the contrary, a *public-question test* (PQT) is one for which the question set can be safely published well before the actual test. The main point here is that it uses a hugely redundant question set, and each candidate is tested on a small subset of the questions, randomly generated by a computer software.

We will show in this article how such a system can be designed without sacrificing reliability and equality. On the other hand, such an test is much more secure (that is, cheating is much harder), synchronization is no longer needed so that candidates can have individual sessions, and the result is quantifiable. Tests can thus be held in much more flexible ways.

In the last section, we will also describe our experiments on public-question tests and their results.

Date: May 2004.

1. The principle of PQT

A *public-question test* is based on a *source* consisting of a large set of questions (question source) as well as a set of rules for randomly selecting questions within the question source.

2. Questions, variations and types

The goal of a test is to evalue the intrinsic competence of the candidates within a certain domain. In practice, the score to a test never reflects exactly the intrinsic capabilities of the candidate. Therefore, one can define the *reliability* of a test system to be the correlation coefficient between the intrinsic competence and the scores to the test. The reliability is always a number smaller than 1. It can be indirectly measured via the correlation between scores of two tests on the same domain and by the same set of candidates.

An important phenomenon that reduces the reliability of a test is that some candidates may be able to get the answers to the questions and memorize them before the test. Under traditional systems, this may occur when the question set leaks, or when people succeeds in guessing the question sets, for example by analyzing the habits of the authors of the questions.

Under a PQT, the question source is public and the selection is random and individual. Here the only problem is to stop people from memorizing the answers. To do so, we notice that memorizing the answer to a question has a cost, in time and effort, to the candidate. And the total cost of memorization will grow at least proportionaly with the size of the question source, unless some shortcut is found.

Now if the size of the question source is sufficiently large so that the memorization cost is significantly greater than the cost to acquire the competence to be evaluated by the test, reasonable candidates will abandon the attempt to memorize, while those who still try to do so will fail due to an unaffordable cost.

Several techniques can be applied to reduce the cost to produce a lage question set. The most important among them is to insert random parameters into a question. As the discussion of these techniques is outside the scope of this article, we simply take some examples to illustrate the situation.

Example 2.1. Computational question. A typical question on statistics gives a series of statistical data and asks the candidate to make analysis on them. If the data are composed of 10 integers, each varying between 1 and 100, the total number of variations is near 10^{20}, well beyond any effort of memorization. However, producing such a variable question does not cost much more than a fixed one.

Example 2.2. Conceptual question based on multiple choice. The statement of such a question can be made to vary, say, via the random variation of a certain number of words. In a simple case where there are ten varying words, each having two possibilities of variation, there are 2^{10} possible variations.

Of course, the answer must depend on the varying words. But it is not necessary that the variation of EACH word changes the good answer. In well-designed questions, knowing on which words the answer depends already requires a good understanding of the concept involved in the question.

Example 2.3. Memory test. This occurs when the test must verify the memorization by the candidate of a large number of small objects, such as the vocabulary of a foreign language. For this case one can easily design questions whose objects are randomly taken from the complete set of objects to test.

The random variation of the questions also makes its memorization more difficult. Instead of simply remembering the answer, now you must also remember every word in the statement, which is much harder. Moreover, the similarity between one question and another is a source of confusion for the human brains, making both harder to remember. The cost of memorizing the answers can thus quickly exceed that of understanding the question and the subject.

What effectively occurs is that candidates will not try to memorize individual answers to such questions with a huge number of variations. They will more likely to *pre-study* them in order to get a better chance at the test. In more developed situations, such pre-studies will most likely become the principal learning activity for acquiring the competence aimed at by the test.

Therefore in practice, the size of the question set is better measured by the number of question types, each question type being composed of a large number of questions that differ from each other only by variations of parameters.

In many cases, a limited number of types are enough to cover whole or part of the goal of the evaluation. Such as examples 2.1 and 2.3. In such cases, it is enough to design a question set that comprehensively covers the corresponding part.

In other cases, each type of questions is a special case of the competence to be evaluated. A significant number of different types will be needed in these cases, so that the whole question set becomes sufficiently representative of the general competence.

The most difficult case is when the creativity and imaginativity of the candidates are to be tested. For this can be only done by putting candidates before situations they have not met

before. One solution is to make the number of types grow to a point that exceeds the memory capacity of the candidate. This is not always possible in practice.

Other solutions exist, especially in many special situations. But it is beyond the scope of this article to discuss them.

3. Equality and variation of difficulty level

In a traditional test, the question set is empirically determined. The need to keep it secret before the test precludes the possibility of making statistical tests on its level of difficulty BEFORE the test. One can only judge the quality of the question set by post-analyzing the results. In practice, it is not rare to see tests whose post-analyzed difficulty level differs significantly from what is intended.

Traditional tests stand because there is the notion of *equality* between candidates: if a question set is difficult, then it is difficult for all candidates, and vice versa. However, this notion of equality is only valid modulo chances. The score of any given candidate will still vary if submitted to two question sets with the same average score over the whole candidates, because for instance he happens to be more familiar with questions in one of them. This "chance factor" is the theoretical upper bound of the reliability of a test, and depends on the technical factors: duration, number of questions, etc.

A PQT system raises the question about equality because question sets are individual and randomly generated. So a particular candidate may have a question set that is easier or harder than another.

Here one must notice that the notion of general level of difficulty of a question set is only statistical. A question set is generally more difficult if the average score by the whole set of candidates is lower. On any particular candidate, such a question set may or may not be more difficult, depending on the above chance factor.

This observation brings us an intuition that if the variation of difficulty levels in the individual question sets are kept below the variation by the chance factor, the first will be absorbed by the second. We will now see the mathematical reason behind this intuition.

To do so, let Q be the set of all possible question sets in a PQT. For any candidate c and any question set $q \in Q$, there is an *expected score* $s_{c,q}$ that c should get if submitted to the question set q. Let the *average expected score* s_c be the average of $s_{c,q}$ for all $q \in Q$. One may consider $s|_c$ to be the intrinsic competence of c. And the standard deviation $\sigma|_c = \sqrt{\frac{1}{|Q|}\sum_{q\in Q}\left(s_{c,q} - s|_c\right)^2}$ is the deviation of the expected score for the candidate c.

Let C be the set of all candidates. The quadratic average of $\sigma|_c$

$$\sigma = \sqrt{\frac{1}{|Q|\,|C|}\sum_{c\in C}\sum_{q\in Q}\left(s_{c,q} - s|_c\right)^2}$$

is called the *total deviation* of the test. This is the global measure of the unreliability of the test.

For any $q \in Q$, let $s|_q = \sum_{c\in C} s_{c,q}$ be the average of $s_{c,q}$ for all $c \in C$. $s|_q$ inversely measures the statistical difficulty level of q.

Now let S be the set of all possible scores and averages. For any value $s \in S$, define $Q_s = \{q \in Q | s|_q = s\}$.Question sets belonging to a same Q_s can thus be considered to be of same difficulty level.

Also for any $s' \in S$, we define $C_{s'} = \{c \in C | s|_c = s'\}$, and define

$$m_{s,s'} = \frac{1}{|Q_s|\,|C_{s'}|}\sum_{q\in Q_s}\sum_{c\in C_{s'}} s_{c,q}$$

to be the average of the expected scores of all $c \in C_{s'}$ for all question sets in Q_s. Note that

$$s' = \sum_{s\in S}\frac{m_{s,s'}}{|Q_s|}\ .$$

And the *level deviation* felt by candidates in $C_{s'}$ is

$$\sigma_{l,s'} = \sqrt{\frac{1}{|Q|}\sum_{s\in S}\sum_{q\in Q_s}\left(m_{s,s'} - s'\right)^2} = \sqrt{\frac{1}{|Q|}\sum_{q\in Q}\left(m_{s|_q,s'} - s'\right)^2}\ .$$

Taking quadratic average over all s', we get the *level deviation* of the test

$$\sigma_l = \sqrt{\frac{1}{|Q|\,|C|}\sum_{c,q}\left(m_{s|_q,s|_c} - s|_c\right)^2}$$

which is the deviation of the statistical difficulty levels in the randomly generated individual question sets.

We may assume that Q is sufficiently big (or that S is sufficiently small with respect to Q), so that $m_{s,s'}$ varies smoothly with s and s'. However, we cannot assume that it does not depend on s', for usually the variation of the difficulty level dosn't have the same effect on candidates with different average expected scores.

On the other hand, for fixed $s, s' \in S$ the standard deviation of expected scores of c on question sets in Q_s is

$$\sigma_{f,s,s'} = \sqrt{\frac{1}{|Q_s|}\sum_{c\in C_{s'}}\sum_{q\in Q_s}\left(s_{c,q} - m_{s,s'}\right)^2}\ ,$$

so that the quadratic average

$$\sigma_{f,s'} = \sqrt{\frac{1}{|Q|}\sum_{q\in Q}\left(s_{c,q} - m_{s|_q,s'}\right)^2}$$

is the deviation of expected scores for $c \in C_{s'}$ without the variation of difficulty level. And we can define the *fixed-level deviation* σ_f of the test by taking the quadratic average over all candidates:

$$\sigma_f = \sqrt{\frac{1}{|Q|\,|C|}\sum_{q,c}\left(s_{c,q} - m_{s|_q,s|_c}\right)^2}\,.$$

This is the "chance factor" as described above, as it measures the deviation of the scores of candidates whith the effect of difficulty level variations extracted.

The following theorem establishes a Pythagorean relation between the total deviation, level deviation and fixed-level deviation.

Theorem. *We have*

$$\sigma = \sqrt{\sigma_f^2 + \sigma_l^2}\,.$$

Proof. By definition,

$$\begin{aligned}
|Q|\,|C|\,\sigma^2 &= \sum_{q,c}\left(s_{c,q} - s|_c\right)^2 \\
&= \sum_{q,c}\left(\left(s_{c,q} - m_{s|_q,s|_c}\right) + \left(m_{s|_q,s|_c} - s|_c\right)\right)^2 \\
&= \sum_{q,c}\left(s_{c,q} - m_{s|_q,s|_c}\right)^2 + 2\sum_{q,c}\left(s_{c,q} - m_{s|_q,s|_c}\right)\left(m_{s|_q,s|_c} - s|_c\right) + \sum_{q,c}\left(m_{s|_q,s|_c} - s|_c\right)^2 \\
&= |Q|\,|C|\left(\sigma_f^2 + \sigma_l^2\right) + 2\sum_{q,c}\left(s_{c,q} - m_{s|_q,s|_c}\right)\left(m_{s|_q,s|_c} - s|_c\right)\,.
\end{aligned}$$

As $m_{s|_q,s|_c}$ is the average of $s_{\bar{c},\bar{q}}$ for $\bar{c} \in C_{s|_c}$ and $\bar{q} \in Q_{s|_q}$, we have

$$\sum_{c\in C_{s'}}\sum_{q\in Q_s}\left(s_{c,q} - m_{s|_q,s|_c}\right)\left(m_{s|_q,s|_c} - s|_c\right) = \left(m_{s|_q,s|_c} - s|_c\right)\sum_{c\in C_{s'}}\sum_{q\in Q_s}\left(s_{c,q} - m_{s|_q,s|_c}\right) = 0$$

for all $s, s' \in S$. Summing up, we get

$$\sum_{q,c}\left(s_{c,q} - m_{s|_q,s|_c}\right)\left(m_{s|_q,s|_c} - s|_c\right) = 0$$

and the theorem follows. □

The meaning of the theorem is that if the level deviation is smaller than the chance factor, its effect quickly becomes negligeable. For example, when $\sigma_l \leq \frac{1}{3}\sigma_f$, we have $\sigma \leq \frac{10}{9}\sigma_f$. This means that the difference between σ and σ_f is impreceptible in practice.

Moreover, in comparison with traditional tests, a PQT allows using statistics data to construct the question source quantitatively. As a result, σ_f can be better bounded and the total deviation smaller than in traditional tests.

It is not the aim of this article to make detailed discussion on the techniques to control σ_l. We take only the simplest example where the test consists only of one "atomic" question, that is, a question on quich a candidate can only "fail" (score 0) or "succeed" (score 1). Also, we suppose that Q contains only two difficulty levels, Q_{s_1} and Q_{s_2}, of equal probability, and that all the candidates have the same average expected score $s' = s|_c$. In this case $s_i = m_{s_i,s'}$, and $s' = \frac{1}{2}(s_1 + s_2)$. We have $\sigma_l = \frac{1}{2}|s_1 - s_2|$.

Example 3.1. Suppose $s' = 0.5$. We have $\sigma = 1/2$, so σ_l can go as high as $1/6$, that is, $\{s_1, s_2\} = \left\{\frac{1}{3}, \frac{2}{3}\right\}$, while remaining imperceptible. This range is fairly easy to respect in practice.

If σ_l grows further to $\frac{1}{4}$, we will have $\sigma_f = \frac{\sqrt{3}}{4}$. One starts to feel the effect of the level deviation, without having the impression of being dominated by it. The point $\sigma_l = \sigma_f$ is reached when $\sigma_l = \frac{1}{\sqrt{8}}$, or roughly $\{s_1, s_2\} = \{0.15, 0.85\}$.

Example 3.2. Suppose $s' = 0.9$. In this case $\sigma = \frac{3}{10}$, and even in the worst case where $\{s_1, s_2\} = \{0.8, 1\}$, the effect of the level deviation ($\frac{1}{10}$) is still negligeable.

And the computation can be easily generated to the case of n atomic questions. Roughly speaking, the effect of level deviation will be neglibeable if for each question, the variation of the difficulty level does not exceed ± 0.15.

4. Concrete experiments

Since year 2002, the author has applied PQT to the final exams of several math courses delivered to first year science students in Université de Nice Sophia-Antipolis, using the software WIMS. The number of total candidates is around 500 each year.

Each time, the question source is published at least 2 weeks before the exam sessions, and software allows simulations under exactly the same technical condition as the real exam. Each exam has several sessions, a student can select one of the sessions. Usually, students have a slight preference for earlier sessions.

The source of each test is composed of a few hundred question types. The software generates a few douzens of questions from the source, devided into 3 groups.

The first group is composed of questions of low difficulty level, and takes around 20% of the score. It has a blocking effect for the rest of the test: candidates can do the rest of the questions only they have achieved sufficient average scores on the first group.

The second group contains questions of average difficulty level, distributed according to the subtopics of the course. It takes around 50% of the score, and has a blocking effect for the third group which contains questions of high difficulty level.

The session is configured to last one hour, but each candidate can try up to 3 sessions within 90 minutes, with the best score being taken into account. This multi-try setup has only a marginal effect on the deviations other than a psychological effect on the candidates. On the other hand, it increases the score spread (that is, the standard deviation on the set of scores by different candidates), by favoring the good ones. In fact, only candidates who can finish a first try ahead of schedule can fully take full profit from a second try. In this case, they usually get higher scores in the second because they are then less stressed.

The multi-try setup also has a side effect, inducing some candidates to abandon the first tries too quickly. Such cndidates often finish with a poor final score. For most of the candidates, this side effect is considerably attenuated by the blocking effects between the question groups. Once a candidate has made efforts to pass through the first group to open the accessibilty of the second, he is more likely to stick to it in order to capitalize these efforts.

Another consequence of the groupwise block effects is to increase the score spread. It also slightly increases the unreliability deviations of the test.

The test sources are also carried over to the same course in the next academic year, with slight modifications or even no modification at all. There is no sign that candidates in the second year are favored from the existence of the same source a year earlier.

We have been prepared for unfairness claims by candidates having received an unusually hard question set. The solution is to propose to cancel their existing scores then give them a new session. (The existing scores must be cancelled in order to stop false claims.) However, up to now no such claims have ever been received. This indicates that in general, candidates do not feel the inequality generated by random question sets.

In our experiences, all the questions are scored automatically by the software, and the candidates can see their scores in real time. However this is not a must for PQT. The software can also randomly generate questions to be scored manually later.

The experiences show that this test method has several important advantages.

(1) First of all, the openness of the source of the final exam is an important mark of confidence shown to the students. The publication of the exam source makes the goal of the entire

course transparent, and students are much more inclined to work hard to reach the goal. Of course, some of them start by taking the simulations as a way to cheat, but they quickly realize that it is the entire course that is behind the random selection of the questions.

(2) Statistics shows a higher reliability. In cases where data are available, we have got correlation coefficients around 0.7 between two tests. While for traditional tests under similar situations, this coefficient usually goes around 0.5 and rarely exceeds 0.6.
The higher reliability also lies in the fact that the global average score, as well as the score spread, can be quanticatively controlled. There is no surprise at the end of the exam to see that it is too hard or too easy, whether in whole or in part, a very frequent phenomenon in traditional tests because the question set is empirical.

(3) The higher reliability results also from the fact that cheating is much more difficult than in traditional tests. Most traditional cheating methods (leak or theft of the questions, copy between neighbors) no longuer apply to PQT. The use of computers can even prevent false identity candidates using biometric authentification.
As a result, the atmosphere of the test room is more relaxe, the surveillance consisting essentially only of preventing candidates from exchanging materials or talking to each other. This in turn helps candidates to proceed with less stress.

(4) It allows more technical flexibility, by eliminating the need for synchronizing among candidates. Each candidates can have an individual test schedule, candidates for different tests can be mixed up in a same test room. Also, a candidate who misses a test due to personal reasons and so on can simply catch up later.
In a more developed situation, a test can be open all year long, waiting for candidates to take it at any time.

(5) As a same test can be used repeatedly over a long period of time or across multiple institutions, statistical data on the scores can be reliably compared.