Journal of Computerized Adaptive Testing

Volume 9 Number 1

October 2022

Improving Precision of CAT Measures

John J. Barnard

DOI 10.7333/2210-0901001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 9, Number 1, October 2022
DOI 10.7333/2210-0901001
ISSN 2165-6592

Improving Precision of CAT Measures

John J. Barnard

EPEC Pty Ltd., Australia

Editor's note: This paper was the author's presidential address at the 2015 conference of the International Association for Computerized Adaptive Testing.

The basic idea of adaptive testing is quite simple and has been implemented for over a century. Dichotomously scored multiple-choice questions are commonly used to obtain response vectors in computerized adaptive tests (CATs). A response is scored as either correct or as incorrect. However, a correct response does not necessarily mean that the examinee knew the answer. Although the standard error increasingly decreases as the provisional ability is estimated, the question is whether the process can be improved at the item response level. In other words, can more information be extracted from a response than a simple 0 or 1? Implementation of option probability theory holds promise to address this question.

Keywords: dichotomously scored items, option probability theory, scoring methods, subjective probability

The goal of measuring properties of an examinee's behavior is to estimate the examinee's level of performance as precisely and efficiently as possible. However, measurement remains imprecise because all measures have some type of error. Sophisticated psychometric models and statistical techniques are still limited by the underlying precision of the measures. When multiple-choice questions (MCQs) are used in measuring instruments, additional challenges such as guessing by examinees need to be dealt with. This issue has been addressed over many decades using corrections for guessing formulae in classical test theory (CTT) and through probabilistic approaches in Rasch measurement theory and in item response theory (IRT). Pseudo-guessing parameters have been used to estimate the amount of possible guessing to account for this and, hence, increase the precision of the measures.

Computerized Adaptive Testing (CAT)

One of the hallmarks of the testing movement has been its long-standing involvement with technology. Only a few decades ago, microcomputers were not as available as they are today, test centers did not exist, software packages were relatively primitive, and the Internet was in its infancy. Computer-administered tests were expensive and limited to a small number of terminals that could be linked to a host mini- or mainframe computer. The development and wider availability of microcomputers accelerated the use of technology in online testing, and today computer-based tests that include clear, high-resolution images, videos, and multimedia are common. The advantages of computer-based assessment have been well-documented.

It was just a matter of time before this mode of assessment was improved with adaptive testing. Computerized adaptive testing (CAT) holds promise to be the most efficient way in which tests can be delivered through tailoring each examinee's assessment to the ability of the specific examinee. Instead of using a fixed-length test, questions are intelligently selected and administered to the examinee to increasingly match the examinee's ability. This means that questions that are either too easy or too difficult are not administered to examinees. Thus, an adaptive test is an interactive process of administering items so that the composition of the test is dynamically modified for each examinee's level of performance.

Two main benefits are offered by CAT over conventional computer-based testing: efficiency and control over measurement precision. Studies have reported reductions in test length of 50% or more compared to conventional testing while maintaining measurement accuracy. A conventional test has items that are either peaked in terms of difficulty, which provides more precise measures within a certain ability range, or distributed over a range, which provides less precision over a wide range of examinee abilities. By not administering items that are off-target for individual examinees, CAT yields high precision over the entire range of examinee abilities. CAT is ideal for discriminating ability precisely and efficiently among people over a broad range of ability levels. The essence of a CAT is not, however, a test that is administered on computer, but rather a test that adapts to the ability of the examinee being tested and thereby improving test efficiency and precision.

Numerous authors have described the components of a CAT:

- A *precalibrated item bank* in which the content of the items is located with item statistics linked to a common interval scale within the framework of a chosen psychometric model.
- A *starting point* that can be random, but the closer the difficulty of the first item is to the ability of the examinee, the fewer items that will be required to arrive at the examinee's final ability estimate with high precision.
- An *item selection algorithm*, which is used to determine which item will be delivered next to best satisfy the selection criteria, provided that limitations such as content and exposure constraints have been met. In general, the algorithm will select a more difficult item if the preceding item has been correctly answered and an easier item if the answer was incorrect. Popular methods include maximum information (Fisher, single point); maximum posterior precision, which uses the largest decrease in variance of the posterior ability distribution; and weighted information, which applies weights from the current posterior ability distribution. A combination of these maximum information and Bayesian methods is often used.
- A *scoring algorithm*, which, after administration of each item, recalculates the examinee's ability measure using all information up to that point. The likelihood function, which measures the goodness of fit of the model used, and an estimator, such as the

(weighted) maximum likelihood estimator, are used to estimate and update the examinee's ability after each response, and the corresponding standard error, which indicates the precision of the measurement, is calculated. Other (Bayesian) methods include the expected a posteriori and maximum a posteriori. The methods can be compared in terms of relative bias and stability.

- A *termination criterion*, which defines the point at which the testing session is stopped. The most commonly used criteria are measurement precision and test length. For variable-length CATs, the standard error is normally used to terminate the test; this implies that the precision of ability estimates will be approximately the same for all examinees although different numbers of items have been administered. For fixed-length CATs, the test is terminated after a specified number of items has been administered. A consequence of this is that the final ability estimates of each examinee will differ in precision.

The question is "Which combination is best for a CAT?"

Measurement Theories and Scoring Methods

Consider unconstrained, unidimensional, dichotomously scored CATs scored item-by-item (assuming that the questions are MCQs). The data for each examinee is then a response vector consisting of 0s and 1s (e.g., 01101111011...).

To analyze the data, it is assumed that the item difficulties are known (previously calibrated) and that the estimates are precise enough to use as true values. In CTT, the item scores are added to obtain a total score that represents the examinee's performance. Criticism of this methodology has been widely published, and the probabilistic methods of Rasch theory and IRT have gained ground over the past number of decades. In Rasch measurement, the probability of a correct response is a function of only the examinee's ability and the item's difficulty:

$$P\left(X_{ij}=1\middle|\theta_j,b_i\right)=\frac{\exp\left[\theta_j-b_i\right]}{1+\exp\left[\theta_j-b_i\right]}. \tag{1}$$

where θ_j is the ability level of person j and b_i is the item difficulty. In the three-parameter logistic IRT model, item discrimination (a_i) and pseudo-guessing (c_i) are added:

$$P(X_i=1|\theta_j,a_i,b,c_i)=c_i+(1-c_i)\frac{\exp\left[\mathrm{D}a_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[\mathrm{D}a_i\left(\theta_j-b_i\right)\right]}. \tag{2}$$

The probability of answering the item correctly (score of 1) or incorrectly (score of 0) is based on the ability of the examinee to select the correct option from a number of options. This might include the examinee's ability to reduce the number of possible correct answers and, hence, increase the probability by using partial knowledge. However, the item score is discrete (i.e., either 0 or 1).

Option Probability Theory

Option probability theory (OPT), a new measurement theory in which the scoring rule is based only on the probability assigned to the correct option, overcomes the issue of discrete scoring. This paradigm is in sharp contrast to Rasch and IRT, in which the probability of selecting the correct option is considered and not the probability assigned to the correct option. The difference is subtle but significant because the probability assigned to the answer considered to be correct is a measure of confidence that the answer is correct rather than a measure of "true" ability.

This approach should not be confused with probability testing, which dates back to the 1960s. In probability testing, the examinee assigns probabilities to all options, and these probabilities are included in the score. Instead of choosing only one option as a response, proportional probabilities are spread over one or more options to indicate the certainty that each particular option is correct. A variety of scoring rules have been suggested, but if probabilities assigned to all options are included in the scoring rules, the item score is influenced by how the probabilities are distributed over the incorrect options.

Consider, for example, the spherical rule $c/\sqrt{\Sigma i^2}$, where c is the probability assigned to the correct option and i represents the probabilities assigned to all the options. Suppose 0.6 (60%) is assigned to the correct option and 0.1, 0.2, and 0.1 to the three incorrect options of a four-choice item. Then the item score is 0.92. However, if probabilities 0, 0, and 0.4 are assigned to the incorrect options, a score of 0.83 is calculated. In both scenarios, 0.6 was assigned to the correct option, but different probabilities were assigned to the incorrect options. Furthermore, note that in the second case it can be assumed that the examinee could rule out two options via partial knowledge and yet the item score is less than for the first case. This is in contrast to the logic that an examinee who could rule out two options through partial knowledge should score higher than an examinee who could not.

Examinees can also manipulate such a scoring system. For example, if an examinee assigns 0.9 to the correct option of seven items and 0.1 to the correct option of three items, the score will be 7.26 (out of 10), while a score of 7 will be obtained for an examinee who assigns 1 to seven items and 0 to three items. Thus, although such scoring rules have a desired property of resulting in a minimum item score of 0 and a maximum item score of 1, the item scores depend on how the probabilities are assigned over the options, which is not desirable.

The item score should be based on only the probability assigned to the correct option and then a logarithmic scoring rule can be used. OPT therefore only considers the probability assigned to the correct option.

- If the option assigned 100% is correct, the examinee will get the maximum possible item score because there is no uncertainty, but if this option is incorrect, a maximum penalty will apply, which suggests a serious fallacy or misconception.
- If an examinee assigns, say, 80% to the correct answer (it does not matter how the remaining 20% is distributed over the remaining options because they are not scored), the examinee will receive a higher score than if 50% was assigned, but the examinee will still be penalized for the 20% uncertainty.
- If an examinee can rule out some options but cannot decide between two remaining options, 0.5 (50%) will likely be assigned to each of these two options. If one of the options is correct, the examinee will receive a score between 0 and 1 but will be penalized for their degree of uncertainty (guessing).
- If an examinee has no idea about the correctness of any of the options, equal probabilities would likely be assigned to each option. This would mean that the examinee does not know the answer and a score of 0 should be assigned.

Only a monotonically increasing logarithmic function can meet these requirements. Because the scoring rule penalizes uncertainty (guessing), the item score will always be less than the maximum possible score as obtained by dichotomous MCQ scoring unless 100% is assigned to the correct option. This is a desired property because the penalty accounts for the portion of uncertainty (guessing). As with Rasch and IRT ability measures, negative and decimal scores can result due to the logarithmic scoring rule, but these can be converted to any convenient reporting scale. To derive a scoring rule, consider Bayes' rule: Assume that each item entails *k* hypotheses and that *i* is correct for each option *k* in the hypothesis H_i. The initial prior hypothesis is that H_i are all equal [i.e., $p(H_i) = 1/k$)]. The conditional probabilities are $H_i \mid M$, where *M* is the mental process and the examinee chooses the option that is correct according to their knowledge and reasoning [i.e., $p(M \mid H_1)$].

According to Bayes' rule:

$$p(H_i \mid M) = \frac{p(H_i) \times p(M \mid H_i)}{p(M)}, \tag{3}$$

and the examinee thus selects the option for which $p(H_i \mid M)$ is a maximum probability as the "correct" option. But this is not necessarily the correct answer, and, therefore, $p(H_i \mid C)$ is a better measure of *M* because it is proportional to $p(M \mid H_i)$, the ability to choose the correct option. Therefore, in the item score the probability assigned to the correct option should be considered and not the option to which the maximum probability is assigned. This is in sharp contrast with dichotomous 0/1 scoring, where an examinee chooses the option most likely correct in their view.

The score should be a monotonic increasing function of $p(H_i \mid M)$ for the correct option *i*, which can be achieved with a linear function of $\ln p(H_i \mid M)$ (i.e., the scoring rule should be a logarithmic rule). The (logarithmic) scoring rule is based on the probability assigned to the correct option only (i.e., $s_i = F(r_c)$ for $0 \le p(i) \le 1$ and $\sum p(i) = 1$). The expected score *E* can be written as a linear function:

$$E = \sum p(H_i \mid M) \times \log p(H_i \mid M). \tag{4}$$

Substitute $p(H_i \mid M)$ with p_i and consider only *r* (for the correct option). To maximize *s*,

$$E = \sum p_i s = \sum p_i F(r_i). \tag{5}$$

Thus, *s* is a function of *F* of r_i and the expected value is $\max \text{iff } r_i = p_i \; \forall i$.

The function *F* is derived through partial differentiation under the condition that $\sum r_i = 1$ using the Lagrange multiplier λ:

$$\frac{\partial \left\{ \left[\sum p_i \times F(r_i) \right] + \lambda \left[1 - \sum r_i \right] \right\}}{\partial r_i} = 0 \;\; \forall i. \tag{6}$$

Thus,

$$p_i \times \frac{\partial F\left(r_i\right)}{\partial r_i} - \lambda = 0\,, \tag{7}$$

yielding

$$\frac{\partial F\left(r_i\right)}{\partial r_i} = \frac{\lambda}{r_i}\,, \tag{8}$$

where $p_i = r_i \ \forall i$ and λ is a constant independent of r_i. Thus, for constants *A* and *B*,

$$F(r_i) = A\ln(r_i) + B\,, \tag{9}$$

where *A* and *B* can be freely chosen.

For scoring, choose *A* and *B* so that

s_i = 0 if $p_k = 1/k$: examinee does not know (a random guess) and

s_i = 1 if r_c = p_k = 1.

Note negative scores for $s < 1/k$. Extreme $s_r = -\infty$ for p_k = 0 to r_c. This can be rectified with a correction parameter or scoring function, and it can be determined how realistic examinees assign probabilities [assign unrealistically high (overestimate) or low (underestimate) probabilities]. If p is assigned to f options of items for which $f\left(t\right)$ are correct, it is expected that $f/f\left(t\right) = r.$ An examinee is well-calibrated if $r\left(j\right) = p\left(j\right)$. Assume a linear relationship *p* = *ar* + *b,* and then *p* can be estimated from *r*. For *k* options, $b = (1-a)/k$ for each item and for every *r* the corresponding *p* is estimated as $\left[ft\left(r\right)/f\left(r\right)\right]$. So for a least squares estimate of *a*, a minimal function of *a* can be chosen so that $F'\left(a\right) = 0$ and *a* can be solved from

$$a = \frac{\sum\left[ft(r)r - \frac{ft(r)}{k} - \frac{f(r)}{k} + \frac{f(r)}{k^2}\right]}{\sum\left[f(r)r^2 - 2f(r)\frac{r}{k} + \frac{f(r)}{k^2}\right]}, \tag{10}$$

so that

$$a = \frac{k\sum\sum t - n}{k\sum\sum r^2 - n}, \tag{11}$$

which yields the best least squares fit if *p* is estimated with

$$p = ar + \frac{1-a}{k}. \tag{12}$$

Note that a more realistic examinee will have a higher score than an extreme examinee.

Conclusions

There is no doubt that modern test theories (Rasch and IRT) have overcome many of the practical problems associated with implementing CTT as the underlying measurement theory in assessment. Likewise, CAT has been demonstrated to be the most efficient mode of assessment and can be compiled to yield precise measures. Although the fit indices of Rasch and pseudo-guessing parameters are significant improvements on the indefensible corrections for pseudo-guessing in CTT, it will never really be known if an examinee guessed or not—unless the examinee is asked. This is realized in OPT, in which the examinee must indicate how sure they are about the correct answer being correct.

In Rasch/IRT, the probability of answering an item correctly is derived from selecting the correct option from a number of options (in MCQs) instead of the probability assigned to the correct option. This subtle difference questions the actual ability trait being measured. Furthermore, the resulting item score is either 0 or 1, which loses information that continuous scores can provide. This can be addressed by using OPT as the underlying model to score items.

Author Address

John J. Barnard, P.O. Box 3147, Doncaster East, 3109, Australia
Email JohnBarnard@bigpond.com