Interpreting Threshold Metrics

Ben O. Smith

May 2021

This document attempts to give the assessment committee guidance on how to interpret threshold knowledge metrics. We will use Monte Carlo simulations to demonstrate the statistical noise given a variety of class sizes, questions, and threshold scores.

Author Information:
Ben O. Smith, Associate Professor
bosmith@unomaha.edu
University of Nebraska - Omaha

P: https://bensresearch.com
D: https://cba.unomaha.edu/econ

Threshold metrics are used extensively for assessment purposes at the University of Nebraska at Omaha. The program-level assessment template encourages programs to use this assessment procedure by asking for a proficiency threshold (hereinafter "threshold") and SLO proficiency target (hereinafter "target"); the General Education assessment template uses similar language. We will refer to the percent of students above the threshold as the "threshold score" for the class/program.

The advantage of this method of assessment is that it is easy to calculate; students are simply categorized as above or below the threshold and a mean is calculated. However, that does not mean they are easy to interpret. Presumably, the goal is to assess knowledge (which is not observable).[1] However, the relationship between the underlying knowledge and the observed threshold score is complex.

To further this discussion, we will define student ability for student i as μ_i (μ would be the average student ability for the population). Similarly we will define q as the number of questions used to assess student i and n as the number of students. T will be the threshold. In principle, we would like to assess μ (average ability), but we only observe the threshold score.

[1] It is important to note that threshold metrics cannot be used to assess student *learning* as there is no measure of existing knowledge before the treatment period.

Assumptions and Monte Carlo Simulations Setup

To assess the statistical noise of the threshold metric, we will use the results of a set of Monte Carlo simulations. In these simulations, the average student ability (μ), number of questions (q), and number of students (n) are specified. The simulation works as follows:

1. Student ability is randomly drawn from a binomial distribution with q questions and average ability (probability answering correct) μ. This determines μ_i for student i.

2. Using μ_i from above, the student answers the assessment questions with μ_i probability of answering each question correct. If student i's average score is equal to or above the threshold T, they are marked as "1", "0" otherwise.

3. The average of the vector of 1s and 0s (for above the threshold or not) is calculated. This is the threshold score for one simulated class of size n.

4. The above steps are repeated 10,000 times and the 10% and 90% quantiles of the generated distribution are extracted.

Given the underlying values and assumptions of the simulation, the university program would expect to observe a value between the 10% and 90% quantiles with 80% confidence (we use an 80% confidence interval as a 95% confidence interval would apply to very few programs with large student enrollments). We will call the difference between the 10% and 90% quantiles "Δ."

The assumptions of the simulations are necessarily strong. Student ability likely does not follow a binomial distribution. However, a binomial distribution is attractive for a few reasons: (1) it is bell-shaped, (2) it is characterized by two parameters – probability of success and number of questions, and (3) it is intuitively justifiable as a student might know or not know a specific bit of knowledge. Nonetheless, the results of these simulations should be used as *guidance*. The specific values generated will vary based on distributional assumptions.

Results of the Monte Carlo Simulation

We have generated Monte Carlo simulations of all combinations of $\mu \in \{0.70, 0.75, 0.80, 0.85, 0.90\}$, $q \in \{5, 10, 15, ..., 100\}$, $n \in \{5, 10, 15, ..., 100\}$, and $T \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. In total, 10,000 Monte Carlo simulations were generated.[2] To use a single simulation as an example, consider $\mu = 0.8$, $q = 25$, $n = 25$, $T = 0.7$. In this Monte Carlo simulation, the 10% quantile of the generated distribution is 0.72 while the 90% quantile 0.92. Thus, the range of the 80% confidence interval is 0.20 (Δ). In this case, the university program would have some confidence in the general region of the results, but should cautiously evaluate any change that could have a detrimental impact on the program. In a more general sense, we can plot all of the simulation results with Δ on the outcome axis and q and n on the two labeled axes.

[2] The complete set of results can be found at http://bit.ly/MonteCarloKR. The simulation source code is available at http://bit.ly/MCKRSim.

Figure 1 shows that as n and q increase, Δ decreases. However, an increase in n reduces Δ at about double the rate that an increase in q reduces Δ. It is also apparent that there are significant gains to increasing n and q when the starting values are low. The gains quickly diminish with values of n or q above 30. Finally, three plots are presented on the figure: orange where $|\mu - T| = 0.0$, blue where $|\mu - T| = 0.1$, and green where $|\mu - T| = 0.2$. When the student

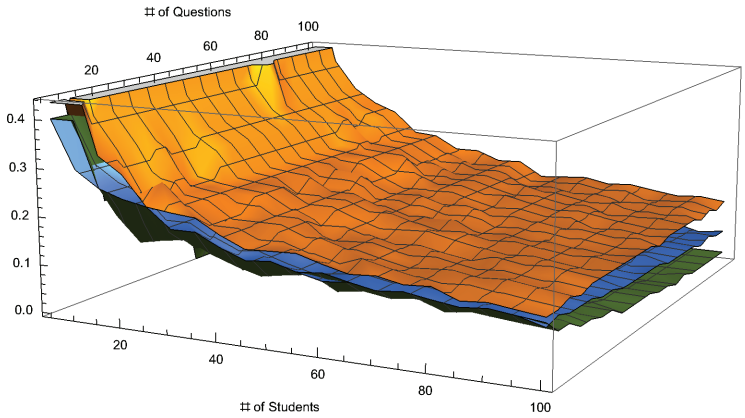

Figure 1: Δ as characterized by n and q. The orange plot is where $|\mu - T| = 0.0$, the blue plot is where $|\mu - T| = 0.1$, and the green plot is where $|\mu - T| = 0.2$.

ability is significantly above the threshold variable the majority of the ability distribution is above the threshold making it less likely that a student will be below the threshold. The net result is the distribution of threshold scores decreases in width.

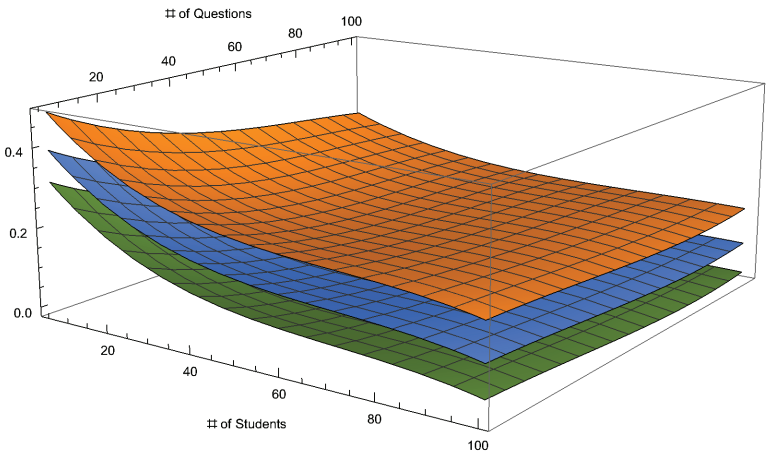

Figure 2: Δ fit to a regression model characterized by n, q, and $|\mu - T|$. The orange plot is where $|\mu - T| = 0.0$, the blue plot is where $|\mu - T| = 0.1$, and the green plot is where $|\mu - T| = 0.2$.

To further this discussion, we have fit the simulation data using Ordinary Least Squares (OLS):

$$\begin{aligned}\Delta &= 0.5703124953376090 \\ &- 0.9373462268502021|\mu - T| - 0.3048565532242980(|\mu - T|)^2 + 3.3921546542560188(|\mu - T|)^3 \\ &- 0.0114899320135041n + 0.0001478718645034n^2 - 0.0000006959575487n^3 \\ &- 0.0066359998152349q + 0.0000662800242221q^2 - 0.0000002727529121q^3 \\ &+ 0.0000639773418393nq - 0.0000000081408744(nq)^2 + 0.0000000000004109(nq)^3 + \varepsilon\end{aligned}$$

Figure 2 plots curves using this equation where $q \in [5, 100]$ and $n \in [5, 100]$. This plot largely matches Figure 1 but it might be more interpretable due to smoother plots.

Guidance

It is apparent that there is a significant level of uncertainly when a university program has less than thirty students and relatively few questions. This statistical noise is exacerbated by the threshold procedure itself. Once a student's average score is determined, their score is converted to either above the threshold (1) or below it (0). Because student scores are reduced to a 1 or 0, this throws information away and increases the level of statistical noise. As a result, more observations are needed when using a threshold metric.

While in an ideal world, all assessment results would be statistically significant, this is an unrealistic expectation for most programs. However, the reduction in statistical noise is not binary. Instead, we can think about noise in assessment data in terms of the risk we are willing to take in the presence of given level of uncertainty. A university program undertaking relatively cost-less and risk-less actions could make those changes with a relatively low level of certainty of the direction of the assessment results. However, if that program wishes to undertake radical changes that might harm student learning in some way, more caution should be exercised.