Yardstick Assessment Strategies
Stratégies en évaluation

Item Response Theory 101: Assessing Candidate Ability

By Dr. Chris Beauchamp

June 2015

In the testing world, mention of Item Response Theory (IRT) conjures up images of cutting-edge state-of-the-art practices needed for any program to be seen as modern and valid.
The purpose of this backgrounder is to provide an introduction to IRT using non-technical language. Since IRT is such a big topic, this blog will focus on how IRT can be used to assess candidate ability. Later backgrounders will focus on other aspects of IRT.

A step back: The purpose of testing
Millions of exams are taken each day. This practice is so common that we lose sight of the assumptions that underlie testing. Unlike physical characteristics such as height and weight (known as "manifest traits") that can be measured directly, characteristics such as technical knowledge, mathematical aptitude, intelligence or artistic creativity (known as "latent traits") must be measured indirectly based on how a person behaves (de Ayala, 2009). In the testing world, how a candidate performs on an exam is used to assess their standing on a latent trait. For example, we conclude that a candidate who scores 95% on an algebra test knows more about algebra than a person who scores 35% on the same test. IRT represents an alternative and more refined way of explaining the relationship between exam performance and a latent trait.

How is candidate ability measured?
There are two major models used in testing and they each measure candidate ability in their own way. The first is Classical Test Theory (CTT) and the second is IRT. CTT has existed for approximately 80 years (Klein, 2005) and is still widely used today. IRT was conceptualized in the 1960s (e.g., Birnbaum, 1968; Lord, 1968) but has not gained prominence until the 1990s when advances in computer technology meant that IRT could be conducted quickly and efficiently.

How is candidate ability measured?

There are two major models used in testing and they each measure candidate ability in their own way. The first is Classical Test Theory (CTT) and the second is IRT. CTT has existed for approximately 80 years (Klein, 2005) and is still widely used today. IRT was conceptualized in the 1960s (e.g., Birnbaum, 1968; Lord, 1968) but has not gained prominence until the 1990s when advances in computer technology meant that IRT could be conducted quickly and efficiently.

An illustration: Classical Test Theory (CTT)

Two candidates each take an identical 10-item test, and score as follows:

Note: ✔ = Correct response; ✖ = Incorrect response

CTT only concerns itself with the total number of correct answers. As a result, CTT would indicate that both candidates are of equal ability. They are both given a score of 60%.

Question	Candidate A	Candidate B
1	✔	✖
2	✔	✖
3	✔	✖
4	✔	✖
5	✔	✔
6	✔	✔
7	✖	✔
8	✖	✔
9	✖	✔
10	✖	✔
Total	6/10	6/10

An illustration: Item Response Theory (IRT)

IRT provides a more sophisticated way of measuring candidate ability. Rather than only being concerned with the number of questions answered correctly, IRT also takes into consideration which questions are answered correctly (similar to a "degree of difficulty" in judged sports such as diving and gymnastics). Let's revisit our previous example.

Question	Candidate A	Candidate B	p-value
1	✔	✖	0.72
2	✔	✖	0.85
3	✔	✖	0.70
4	✔	✖	0.75
5	✔	✔	0.79
6	✔	✔	0.70
7	✖	✔	0.71
8	✖	✔	0.65
9	✖	✔	0.66
10	✖	✔	0.40
Total	6/10	6/10	

Note: p-value refers to the proportion of candidates who responded to the question correct (e.g., p=0.75 means that 75% of candidates responded correctly);
✔ = Correct response; ✖ = Incorrect response

Although both candidates scored 6 out of 10, Candidate B correctly answered more difficult questions (avg. p-value = 0.65) than Candidate A (avg. p-value = 0.75. As a result, an IRT analysis would conclude that Candidate B is stronger than Candidate A.

IRT also differs in how candidate ability is reported. Rather than using raw exam scores derived from summing up correct answers, IRT uses the Greek letter theta (Θ) to describe candidate ability. Theta scores range from -3 to +3. Candidates with a Θ of zero are average candidates. Candidates with negative Θ scores are below average and candidates with positive Θ scores are above average. Using the example above, Candidate A may have a Θ score of 0.5 while Candidate B may have a Θ score of 1.0.

The pros and cons of IRT

IRT represents a more sophisticated and computationally complex way of quantifying candidate ability. However, there are a number of pros and cons of IRT that should be considered.

Pros

- IRT is less influenced by the properties of the test and gives a more accurate picture of a candidate's ability. For example, if a candidate obtained a score of 60% on one test, and two months later, obtained a score of 85% on a similar test, did the candidate improve or was the second test simply easier.
- There have been a number of exciting advances in testing over the past 20 years. Many of these advances rely on IRT and cannot be supported by CTT (e.g., Computer Adaptive Testing (CAT)). Without IRT, some of these advanced techniques are not available.

Cons

- IRT requires some very stringent statistical assumptions. These are often not met, and if these assumptions are not met, results from an IRT analysis may be invalid.
- The implementation of IRT can change the entire exam development and scoring process. This is time consuming and expensive. The benefits may not be worth the investment.
- IRT requires large candidate numbers in order for it to be effective. At minimum, 250+ candidates are needed in order to calculate IRT statistics. Many exam programs simply do not have enough candidates to support IRT.

What's next?

As indicated earlier, IRT is a very big topic. Future blog entries will focus on other aspects of IRT such as determining item parameters and test characteristics, and how IRT can be used to conduct equating and item bank calibration.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

De Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. New York, NY: The Guilford Press.

Kline, T.J.B. (2005). Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses. In T.J.B. Kline, Psychological Testing, A Practical Approach to Design and Evaluation. Los Angeles, CA: Sage.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistical model. Educational and Psychological Measurement, 28, 805-813.