

Ability Estimation with Item Response Theory

White Paper

Nathan A. Thompson, Ph.D.
Vice President, ASC
Adjunct Faculty, University of Cincinnati

Contact Information

Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, Minnesota 55114
Voice: (651) 647-9220
Fax: (651) 647-0412
E-Mail: solutions@assess.com
www.assess.com

Copyright 2009, Assessment Systems Corporation

Table of Contents

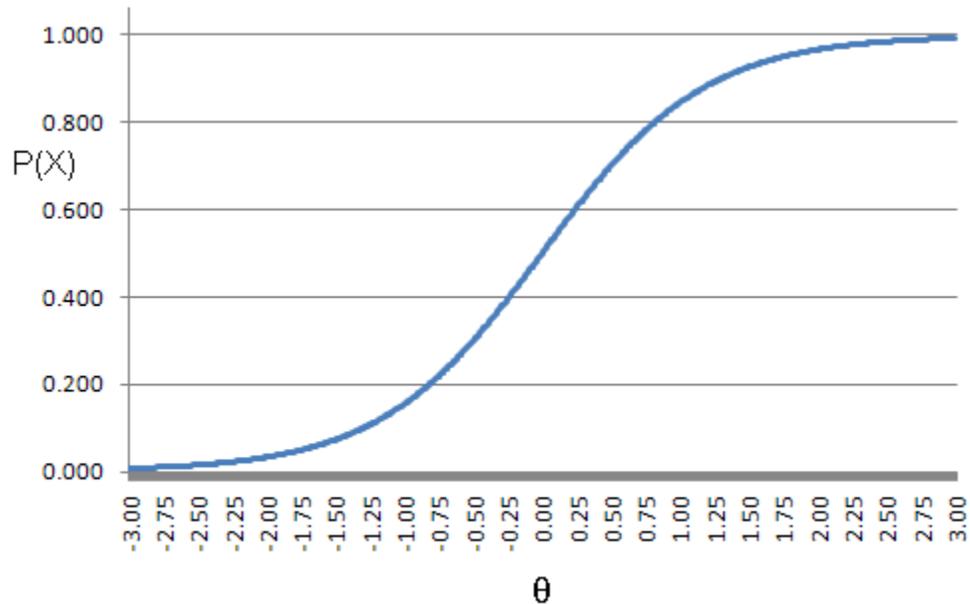
Introduction.....	1
The likelihood function.....	2
Finding the maximum.....	5
Newton-Raphson method.....	6
Possible issues in estimation.....	9
Conclusions.....	11

Introduction

Item response theory (IRT) is a psychometric paradigm for the construction, scoring, and analysis of test forms and items. It offers several advantages over its predecessor, classical test theory, due in part to its greater sophistication. However, this same sophistication can be perceived as a disadvantage due to the greater level of mathematical complexity. One example of this is the scoring of examinees. With IRT, examinees are no longer scored by the number of items they answer correctly. Instead, IRT provides a precise estimate of their location on the underlying distribution of ability, analogous to the standard z scale. The purpose of this paper is to provide a *graphical* introduction to the process of ability estimation with IRT.

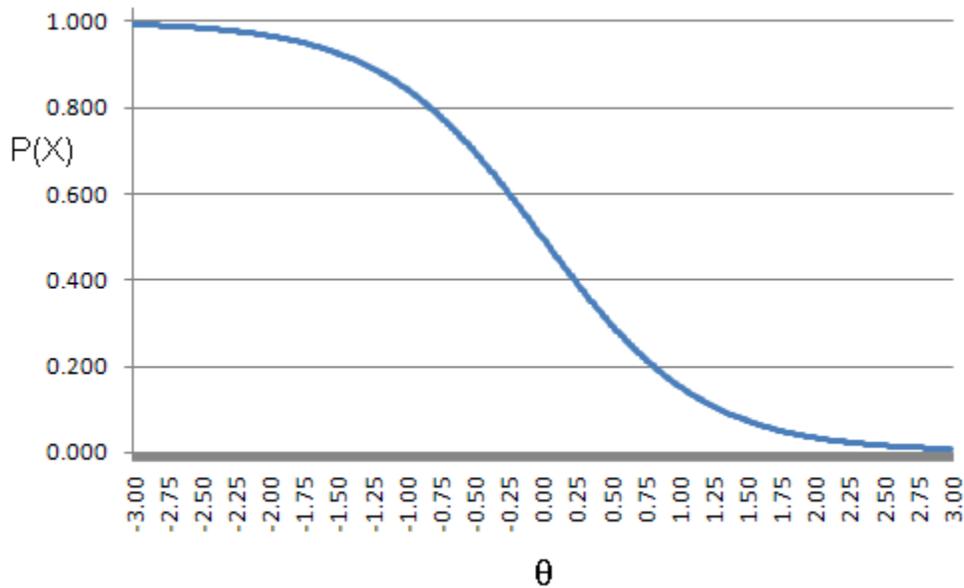
The basic premise of IRT is that the probability of a correct or keyed response is a function of an underlying trait or ability, denoted by the Greek letter theta (θ), with a scale typically depicted as ranging from -3 to 3, with 0.0 representing average ability. Figure 1 presents an example of this function, called an *item response function*. As the trait level increases from left to right, the probability of a correct response $P(X)$ increases.

Figure 1: Example IRF



The probability of an incorrect response is represented by the complement of this function, $1 - P(X)$, as shown in Figure 2.

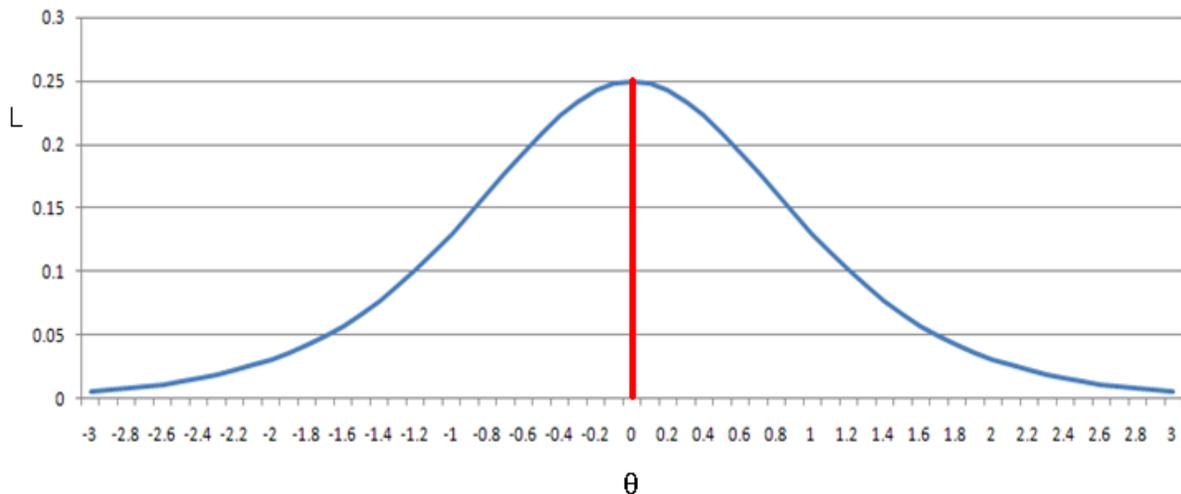
Figure 2: Example of $1 - P(X)$



The likelihood function

To score examinees with IRT, these functions – the IRF for a correct response and the complement for an incorrect response – are multiplied to obtain a single function referred to as the *likelihood function* (LF). This presents the likelihood of each θ level given the responses of an examinee. For example, if an examinee has one correct answer and one incorrect answer to the example item shown in Figures 1 and 2, these curves are multiplied across θ to produce the example found in Figure 3.

Figure 3: Example likelihood function



The likelihood function is defined as

$$L \approx \prod_{i=1}^n P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \quad (1)$$

where \mathbf{u} is a *response vector* of 1s and 0s of person j to items i numbered 1 through n . Note what this does to the exponents; a correct response denoted by $u_{ij} = 1$ means that P remains while Q drops out because any number to the $(1-1) = 0$ power is 1.0. The opposite happens for an incorrect response denoted by $u_{ij} = 0$. This is how the idea of multiplying IRFs and their compliments is operationalized.

If both correct and incorrect responses are present, the likelihood function will tail off to 0.0 towards the extremes, as seen in Figure 3. The location of the likelihood function, whether it is centered on 0.0 or somewhere else on the x -axis, is a function of both the items that are presented to an examinee and the responses given. A different location of the likelihood function leads to a different location of the maximum; this is how IRT provides precise θ estimates.

Some more detail regarding the figures will help explain this. The items in Figures 1 and 2 have a location parameter called b_i of 0.0, as the middle of the curve for each is located at 0.0. If an examinee gets one item correct and one item incorrect at this location, then it makes sense that the examinee θ estimate (similar to location) is also 0.0, as seen in Figure 3. If the two items had a location of -1.0, then the examinee would also have a θ estimate of -1.0. If the two items had a location of 1.32, then the examinee would have a θ estimate of 1.32. This becomes more complex as more items are administered, each of which will probably have a different location, but the concept of determining the maximum in Figure 3 remains the same.

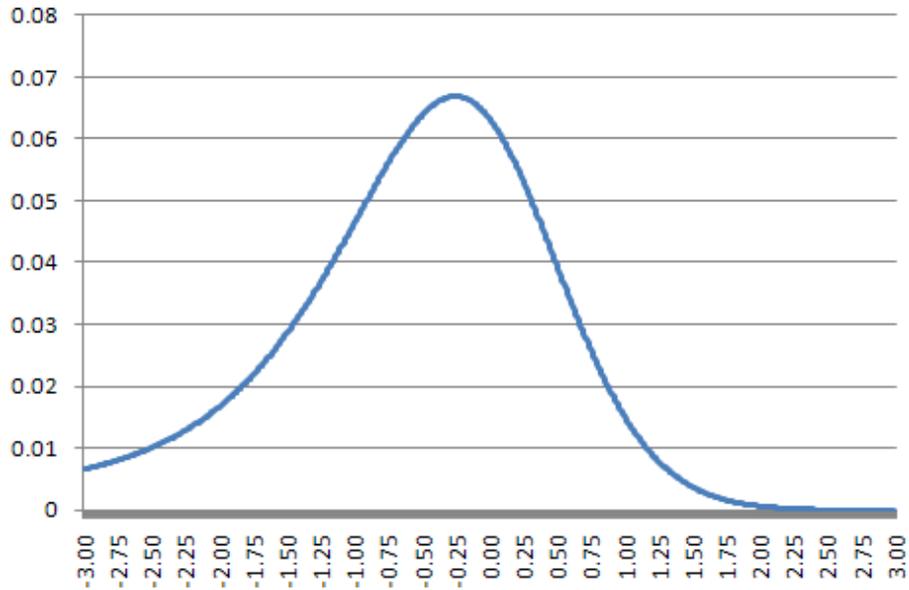
The likelihood function is not only used in θ estimation. It also provides an index of the error involved in estimating each person's θ level, which is the *standard error of measurement (SEM)*. Whereas θ estimation operates by determining the middle of the curve, SEM is based on the spread of the curve. A higher SEM indicates that the curve is more spread out, and therefore the score is less precise.

There are three approaches commonly used to estimate θ within this framework. The predominant method of θ estimation is called *maximum likelihood*, called so because it simply finds the highest point on the likelihood function and returns the value of θ at which it occurs. This is marked by the red line in Figure 3; the θ estimate is 0.0. A common variant of this is the Bayesian modal estimation procedure, also called *maximum a posteriori*, or MAP, where this likelihood function is multiplied by an additional curve that represents an assumed population distribution. This approach was utilized in Owen's (1975) Bayesian adaptive testing algorithm. A further variant is to take this Bayesian-modified curve and rather than find the maximum point, find the average value as weighted by the function. This is referred to as Bayesian expectation *a posteriori* (EAP).

The primary advantage to the Bayesian methods is that they work when an examinee has only correct or only incorrect responses, which is called a *nonmixed response vector*. Suppose that an examinee has answered two items, both of them correctly. Multiplying two IRFs like the curve in Figure 1 only produces a similarly shaped curve, one that increases indefinitely as a function of θ . In that case, no maximum exists, so maximum likelihood is not possible. Multiplying the curve by a bell-shaped curve will give this function a maximum.

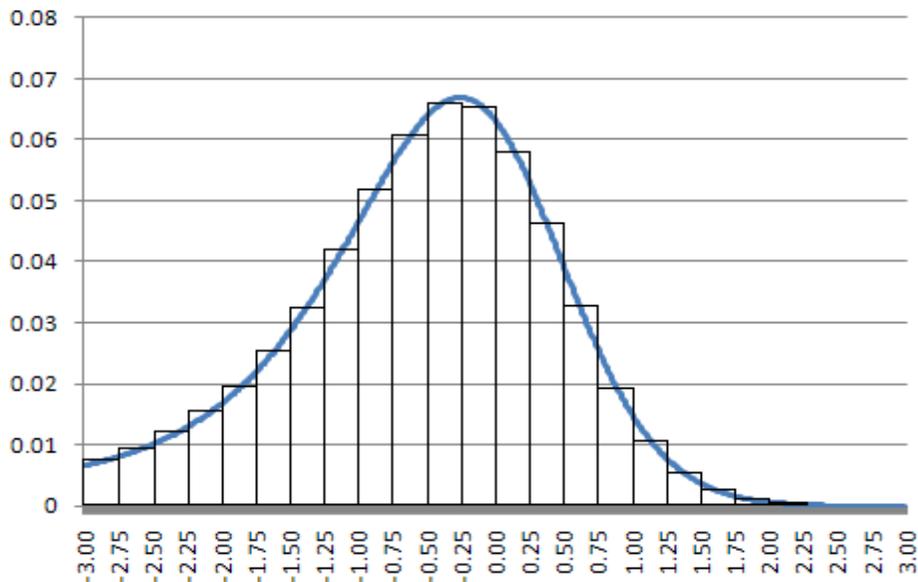
The Bayesian EAP method addresses a further issue: asymmetrical likelihood functions. The likelihood function in Figure 4 represents a typical shape found with 3PL items. Because of the lower asymptote on the left-hand side, the likelihood function can be higher there than on the right-hand side. Finding the simple maximum of the curve does not take this uneven distribution into account.

Figure 4: Asymmetrical likelihood function



Instead, the EAP method finds the average of the curve (rather than the mode) by dividing it up into slices, similar to the Riemann integration of a curve. The average of the values is then calculated. This is demonstrated in the following figure, where more weight would be given to the slices on the left-hand side of the curve, leading to a EAP estimate that is slightly less than the MAP estimate.

Figure 5: Bayesian EAP depiction



So if the Bayesian methods address important issues, why are they not always preferred over the MLE? The reason is that the Bayesian methods are biased. By multiplying the likelihood function by another curve, a bit of arbitrariness is added, and if the Bayesian prior is the standard normal curve – which it often is – the θ estimates will be biased inward to a small extent. The observed standard deviation of final θ estimates might be 0.94 instead of the 1.0 assumed by the distribution.

Finding the maximum

One of the most important algorithms in the application of IRT is that which is utilized to determine the maximum of the likelihood function. This is a necessary component of both the MLE and Bayesian MAP methods; it is not utilized in Bayesian EAP. The most straightforward approach would be to evaluate the value of the likelihood function for every value of θ in the desired range to the desired level of precision. For example, if the desired level of precision is 0.001, the value of the likelihood function could be calculated from -4 to 4 at each point and the highest point recorded. This is called the *brute force* method. However, this is computationally intensive, as it entails a total of 8,000 calculations.

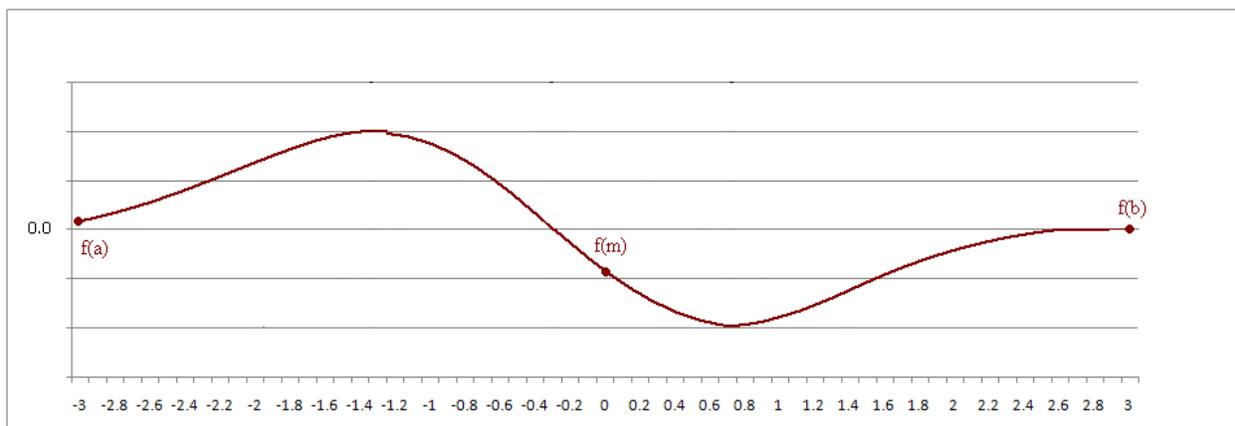
Alternatively, an iterative algorithm could be applied to find the maximum with fewer required calculations. Two iterative methods that are useful for this process are the *bisection method* and the *Newton-Raphson method*. The bisection method is simpler, but not as efficient, and therefore the Newton-Raphson method serves as the primary iterative estimation algorithm.

Bisection method

The bisection method operates, as would be expected, by dividing a range of the LF into two sections and evaluating the midpoint. It was originally developed to find the root of any function, namely $f(x)=0$. Because the maximum of a function is where the derivative is equal to 0, the bisection method lends itself well to evaluating the derivative of the LF, but it can also be adapted to the LF itself.

For the derivative, the bisection method operates as follows. For the first round of the iteration, a wide range is specified from a lower bound a to an upper bound b , such as -3 to +3. As long as the root of the derivative is between these points (the reason a wide range is needed), the value of the derivative at a will be positive and the value at b will be negative (Figure 6).

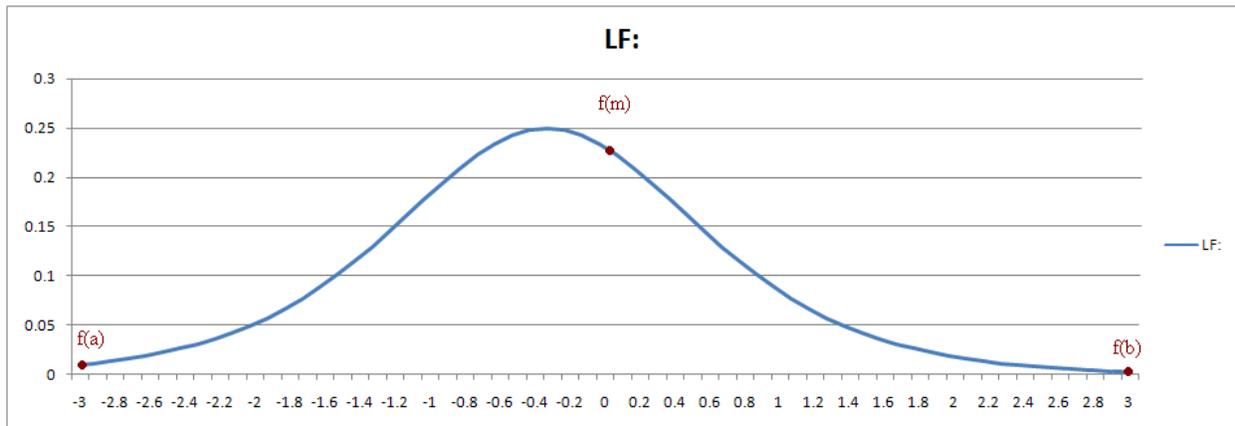
Figure 6: Bisection method on the LF derivative



Next, the midpoint m is found, which is 0 in the example. The derivative is evaluated at this point, and if the value is negative, this implies that the root is between a and m , because $f(a)$ is positive and $f(m)$ is negative. If the value is positive, it implies that the root is between m and b . Based on this knowledge, we can update the range. In the example, the value is negative, and therefore we can eliminate the possibility that the root is between m and b . The midpoint m is then reassigned as the upper bound for the next iteration, and the process is repeated. In the next iteration, we can see that the value of the derivative at the new midpoint -1.5 is positive, which means that -1.5 will be reassigned as the lower bound for the third iteration. This process continues to a user-designated level of accuracy, such as 0.01.

The same type of iteration can be applied to the LF itself, as shown in Figure 7. If the value of $f(a)$ is greater than the value of $f(b)$, this implies that the maximum is between a and m . The midpoint is again reassigned as the new upper bound, and the second round evaluates -3 to 0 with a midpoint of 1.5. Because $f(0) > f(-3)$, the midpoint is reassigned as the lower bound for the third iteration. Note that this is operating completely parallel to the bisection method applied to the derivative. A further adaptation to make it more efficient would be to evaluate the value of the function at two points near to m , such as $m + 0.1$ and $m - 0.1$, and the section with the lower value is eliminated. This approach would eliminate the problems due to asymmetry discussed below.

Figure 7: Bisection method on the LF

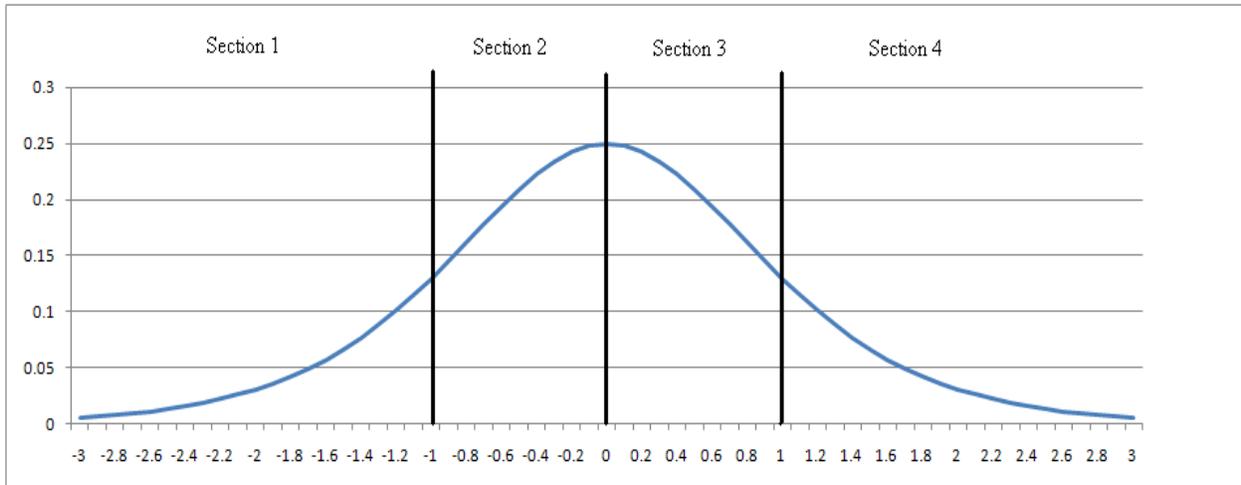


Newton-Raphson method

The iterative procedure that is often used to obtain maximum likelihood θ estimates is the Newton-Raphson procedure. The basic principle of this procedure is that it evaluates the first and second derivatives at a point on θ that is the current iterative estimate, and these values can tell the procedure where on the LF the point is relative to the maximum. The iteration then adjusts accordingly. For example, the values can tell the procedure that the current estimate is only slightly below the maximum, and then the procedure adjusts upward by a small increment. It is therefore more efficient than the bisection method because the increments are variable rather than always being halves.

The Newton-Raphson procedure is best explained graphically by dividing the LF into four sections. First, draw a vertical line at the maximum. Next draw vertical lines at each of the inflection points. This will produce a figure like Figure 1.

Figure 8: Example LF



The sections reflect what is occurring in terms of the slope of the function:

- Section 1: Slope is increasing at an increasing rate**
- Section 2: Slope is increasing at a decreasing rate**
- Section 3: Slope is decreasing at an increasing rate**
- Section 4: Slope is decreasing at a decreasing rate.**

The first and second derivatives of the LF also reflect these sections, as seen in Figures 2 and 3.

Figure 9: First derivative of example LF

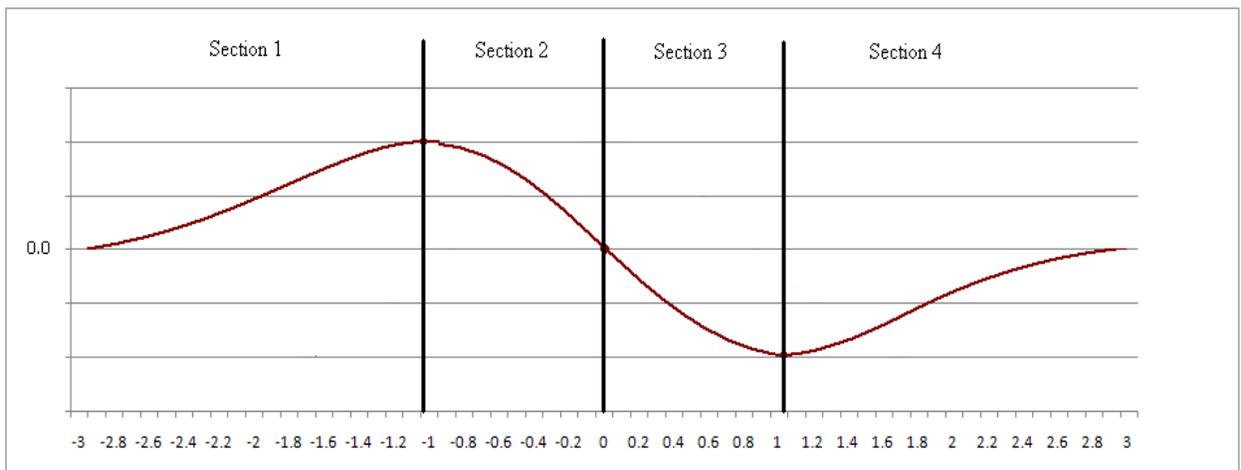
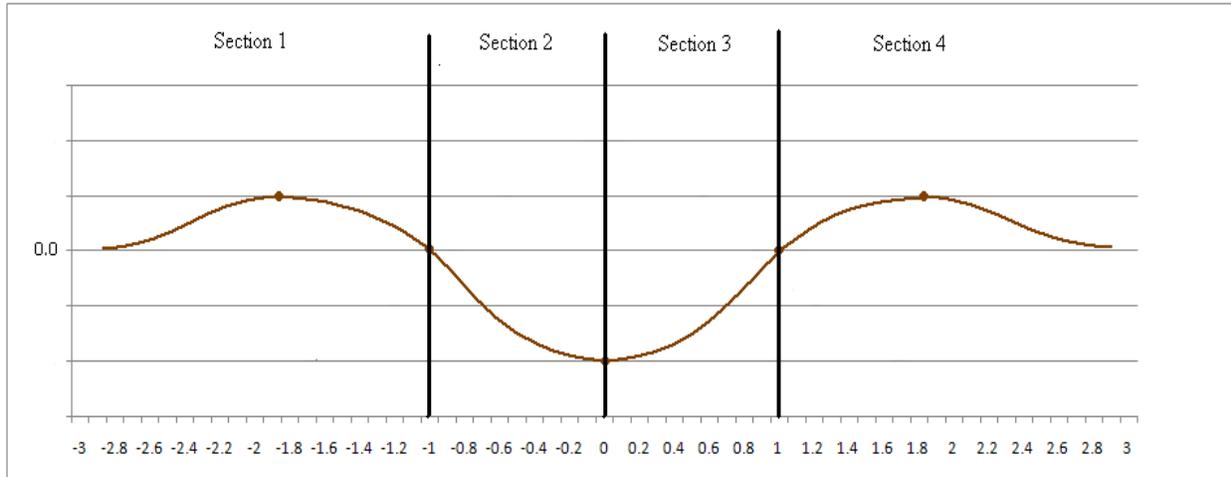


Figure 10: Second derivative of example LF



The basic form of the Newton-Raphson iterative adjustment is $f'(x)/f''(x)$, or the first derivative divided by the second derivative. The four sections as presented above translate into four types of results for the Newton-Raphson with regards to the sign of the values:

- Section 1: Positive/Positive = Positive result**
- Section 2: Positive/Negative = Negative result**
- Section 3: Negative/Negative = Positive result**
- Section 4: Negative/Positive = Negative result.**

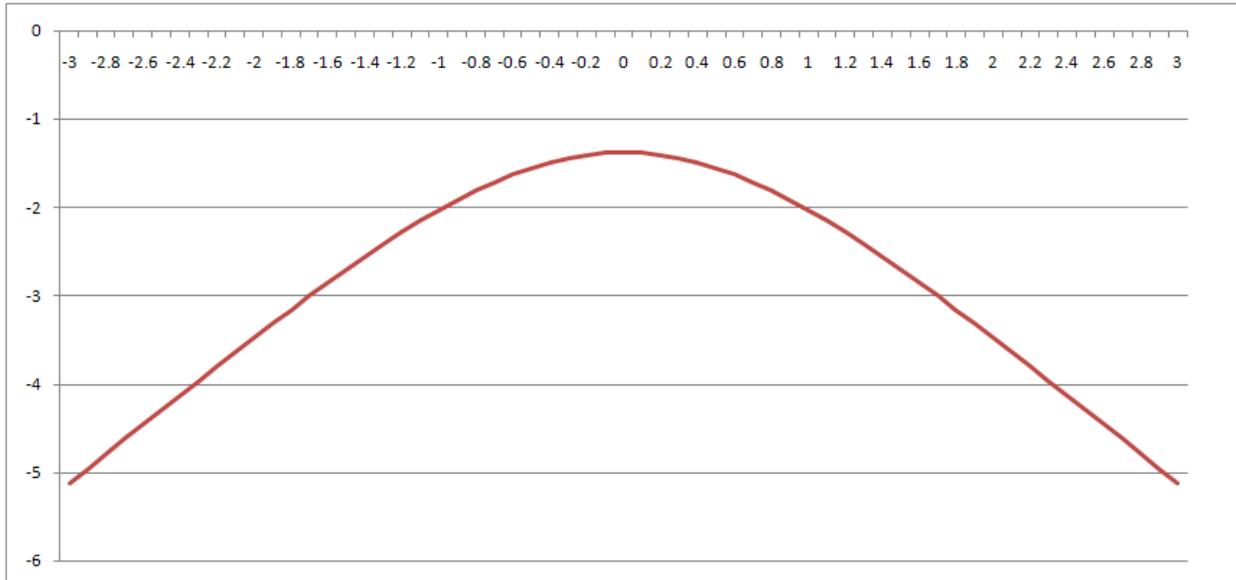
If the current iterative estimate is in Section 1 or Section 2, then the adjustment must be in a positive direction. The adjustment should be in a negative direction for Sections 3 and 4. This occurs naturally when the estimate is in Sections 1 and 4, but not in Sections 2 and 3.

Therefore, a modification must be implemented to ensure that the adjustments are in the correct direction. Given the results above, it is evident that one such modification is to assess the absolute value of the second derivative in the denominator. This makes the denominator positive, which in turn provides a positive result in Section 2 and a negative result in Section 3.

The same issue is present when utilizing the log-likelihood function rather than the likelihood function. However, the log of the LF has no inflection points (Figure 11), so it only has what are equivalent to Sections 2 and 3. As these are the two sections that require a modification of the sign, a similar constraint must be placed on Newton-Raphson estimations of the maximum of the log-likelihood function.

An additional constraint that is often useful is to restrict the *magnitude* of the adjustment values. The reason for this is that the second derivative crosses the x -axis twice and also asymptotes toward the x -axis twice. In these situations, a very small value of the second derivative will be produced along with a relatively large value of the first derivative (see the values -1.0 and 1.0 in the figures). Dividing a relatively large number by a very, very small number produces a very, very large number, which can send the iteration very far from where it needs to be. Therefore, it is prudent to restrict the size of the iteration step to a modest value, such as only 0.5 or 1.0.

Figure 11: Natural log of the LF



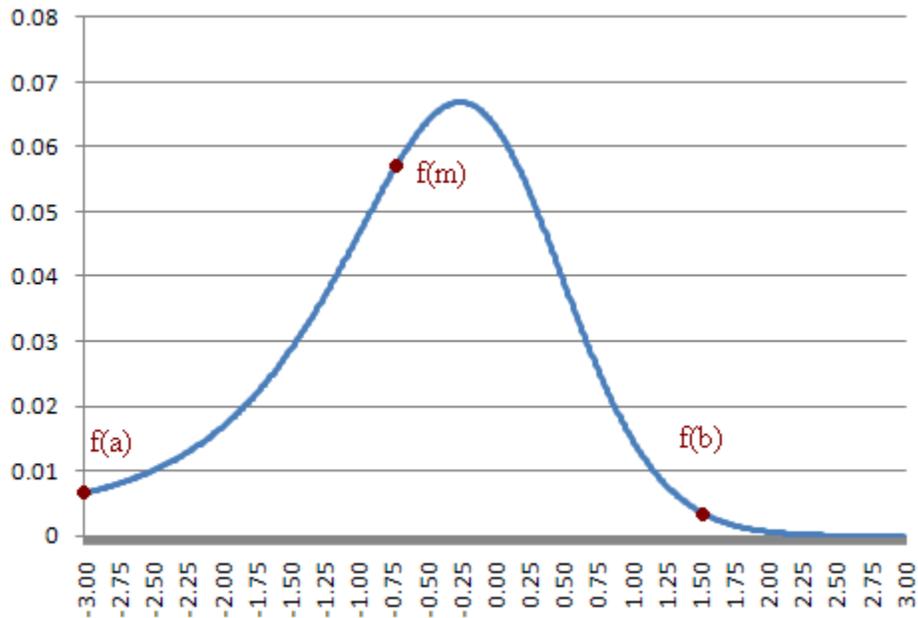
A major benefit of the Newton-Raphson method is that it provides a basis for an estimate of the observed standard error of measurement (SEM) using the second derivative. However, this can be calculated independently, and therefore can still be used if θ is estimated with the bisection or brute force methods.

Possible issues in estimation

As previously discussed, the brute force method is quite precise because it evaluates all possible values, but is extremely inefficient for the same reason. This is the primary reason for the development of the bisection and Newton-Raphson methods for θ estimation decades ago, when computers were capable of only a tiny fraction of the speed possible today. However, given the speed of computers in the 21st century – which number in the millions of instructions per second – this inefficiency might translate to a negligible difference in calculation time.

The iterative methods are very fast, but have the disadvantage that they can get “hung up” on issues that directly affect the logic of the iteration. For instance, the bisection method, because it relies on symmetrical divisions of the LF, is susceptible to problems when the LF is asymmetrical. As mentioned earlier, this is often the case when the 3PL is used because the lower asymptote leads to higher values on the left-hand side of the LF. This is shown in Figure 12. Suppose that $a = -3$ and $b = 1.5$. Because $f(a) > f(b)$, the bisection method would eliminate everything to the right of m – but the maximum (-0.25) is contained in that section. If the bisection method continued, the MLE would be -0.75.

Figure 12: Asymmetrical likelihood function with inaccurate bisection



Note, however, that this problem only occurs when the actual LF or its log is utilized. If the derivative is used, this is not an issue. The derivative at m is still positive, so that the left-hand side would be correctly eliminated.

One problem that affects both methods is equal values on either side. In the bisection method with the LF, this occurs when $f(a) = f(b)$, in which case the algorithm cannot choose which section to eliminate. With the derivative, this is manifested as $f(m) = 0$, which is actually beneficial because then the MLE is found and the iteration can end early. This presents another reason to utilize the derivative rather than the actual LF or its log.

A similar problem can occur with the Newton-Raphson method. If the adjustment moves the next iteration to a point on the other side of the curve with the same values of the derivative and second derivative, the adjustment at this next round will be the same value, sending the subsequent round back to the original point. The method will oscillate indefinitely between the two points. To help prevent this, some type of unequal bounding can be implemented.

An additional problem that affects both iterative methods is the issue of local maxima. In rare cases, the likelihood function is not strictly increasing on the left side and strictly decreasing on the right side. There can be a small bump on either side where the derivative is zero, switches signs, and then switches again. If one of the iterative methods finds itself in the range of this local maximum during the early rounds, it can get caught. Because the iterative methods ignore the rest of the likelihood function, it can never realize that there is a higher maximum in a different range of θ .

Conclusions

In summary, there were seven methods discussed for θ estimation:

1. MLE, brute force
2. MLE, bisection
3. MLE, Newton-Raphson
4. MAP, brute force
5. MAP, bisection
6. MAP, Newton-Raphson
7. EAP.

This leads to the issue of which one to use, or which set to use in conjunction. Because each method has advantages and disadvantages, it is often advantageous to utilize more than one method in software for θ estimation. If the Newton-Raphson method gets hung up on a local maximum, the only way to know so is to compare it to the results of another method. Of course, because the bisection method is also iterative, it could also have gotten hung up on the same local maximum.

The brute force and EAP methods have the distinct advantage of not being susceptible to iterative issues simply because they are not iterative. Therefore, one of these should be utilized as the comparison standard for the iterative methods. This requires much more computation, but also provides the necessary measure of quality assurance. Of course, this raises an important question: if a brute force method is used, which evaluates all possible θ estimates to a desired level of accuracy, is it even necessary to employ one of the shortcut iterative methods? Such discussions are an important aspect of developing software for θ estimation.

ASC is the leading source for software regarding IRT and computerized adaptive testing (CAT). Xcalibre 4 is the most user-friendly and comprehensive IRT analysis software available: a free trial copy is available at <http://www.assess.com/xcart/product.php?productid=569>.