Item Response Theory and Fisher Information for Small Tests

Bivin Philip Sadler and S. Lynne Stokes

Abstract Item response theory (IRT) is a comprehensive paradigm for modeling test performance on the item level in contrast to the more general test-level assessment of classical test theory (CTT). Given the added flexibility provided by item-level modeling, IRT has become the predominant theory used in high-stakes tests such as the SAT, LSAT, and GRE. IRT not only provides an estimate of the examinee's ability but also describes methods to estimate the variance (in terms of Fisher Information $I = 1/Var(\hat{\theta})$) of the ability estimate. As will be explained and demonstrated in this chapter, however, these methods are asymptotic and are inadequate for smaller tests with 15 or fewer questions (as might be found in a computer adaptive test). In addition to illustrating the difference between the IRT estimate and the true variance of the ability estimate for smaller tests, an alternative method of variance estimation will be provided and demonstrated.

1 Basics

Although IRT provides a powerful model in which to design and assess tests, its fundamentals are simple. For each item, the probability of a correct response is modeled with a logistic curve (Fig. 1a) in which the x-axis represents the ability range from −3 to 3 and the y-axis represents the probability of a correct response. The curve is known as an item characteristic curve (ICC). The two-parameter logistic version of the model (known as 2PL) describes the probability of a correct response as

B. P. Sadler (✉)
Master of Science in Data Science, Southern Methodist University, Dallas, TX, USA
e-mail: bsadler@mail.smu.edu

S. L. Stokes
Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
e-mail: slstokes@smu.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_12

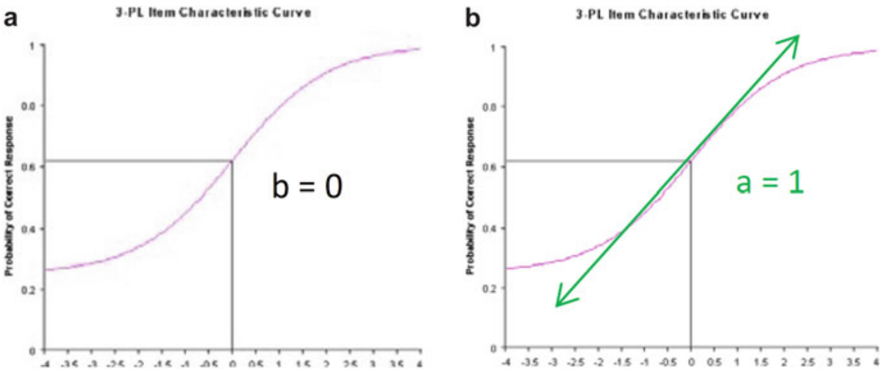


Fig. 1 (**a**) Item characteristic curve (ICC) with difficulty $b = 0$; (**b**) Same ICC showing that the discrimination for the item is $a = 1$

$$p_i(\theta) = \frac{1}{1 + e^{-1.702 a_i(\theta - b_i)}}. \tag{1}$$

The parameter b describes the item's difficulty. Specifically, it is the point on the x-axis where the examinee has probability 0.5 to answer the item correctly (Fig. 1a). The parameter a is the discrimination parameter, which represents the slope of the ICC at b. It describes how well the item ascertains the examinee's ability above or below the difficulty of the item (Fig. 1b).

There are other forms of the IRT model for items. Among these are the one-parameter Rasch model, which retains the difficulty parameter but sets $a = 1$. Another version is the three-parameter logistic (known as 3PL) model, which is often used for multiple-choice items, because it includes a guessing parameter. In this chapter, we illustrate our methods with the 2PL IRT model as defined in (1).

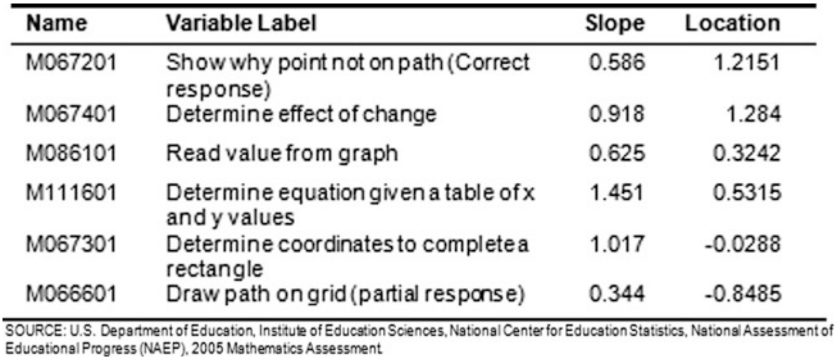
2 Estimation

The IRT model can be used to provide an estimate of the examinee's ability from their responses, when the item parameters are known. If the item parameters are unknown, they can be estimated simultaneously with the ability measures from a sample of examinee responses. For simplicity we assume that the item parameters are known and focus on estimation of ability only.

Maximum likelihood estimation of ability is illustrated with the data from the 2005 National Assessment of Educational Progress (NAEP) Math Assessment. Table 1 displays the slope (a) and location (b) parameters for six actual sample items from the NAEP test (Beaton et al., 2011).

Table 2 shows responses to these items from four fictitious examinees (Beaton et al., 2011). Let z_i denote the indicator of a correct response, i.e.,

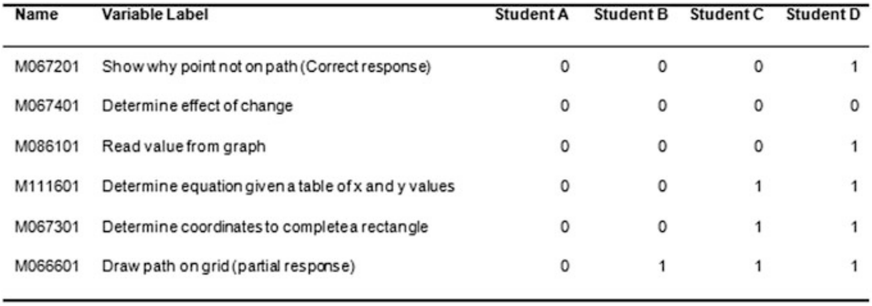
Table 1 Item parameters for the six items referred to in Table 2



Name	Variable Label	Slope	Location
M067201	Show why point not on path (Correct response)	0.586	1.2151
M067401	Determine effect of change	0.918	1.284
M086101	Read value from graph	0.625	0.3242
M111601	Determine equation given a table of x and y values	1.451	0.5315
M067301	Determine coordinates to complete a rectangle	1.017	-0.0288
M066601	Draw path on grid (partial response)	0.344	-0.8485

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment

Table 2 Four different students' responses to six different math questions. A correct response is indicated by a "1" and an incorrect response by a "0"



Name	Variable Label	Student A	Student B	Student C	Student D
M067201	Show why point not on path (Correct response)	0	0	0	1
M067401	Determine effect of change	0	0	0	0
M086101	Read value from graph	0	0	0	1
M111601	Determine equation given a table of x and y values	0	0	1	1
M067301	Determine coordinates to complete a rectangle	0	0	1	1
M066601	Draw path on grid (partial response)	0	1	1	1

$$z_i = \begin{cases} 0, & \text{incorrect response to item } i, \\ 1, & \text{correct response to item } i. \end{cases}$$

As an example, Student C answered the first three questions incorrectly and the last three correctly. If the six item responses are independent, the likelihood of Student C's ability given their observed pattern of responses is seen from (1) to be

$$L(\theta|Z) = \prod_{i=1}^{6} \left(\frac{1}{1+e^{-1.702a_i(\theta-b_i)}} \right)^{z_i} \left(1 - \frac{1}{1+e^{-1.702a_i(\theta-b_i)}} \right)^{1-z_i}.$$

Student C's likelihood $L(\theta|Z)$ is shown as the bold curve in Fig. 2. The thinner curves show the item characteristic curves of the six items composing the test. Ability is measured on the same scale as the location parameter. On this NAEP test, the range of ability is -3 to 3, with a mean of 0. An iterative Newton-Raphson-type procedure is usually used to maximize this likelihood function to determine the maximum likelihood estimate (MLE) of Student C's ability. Visual inspection

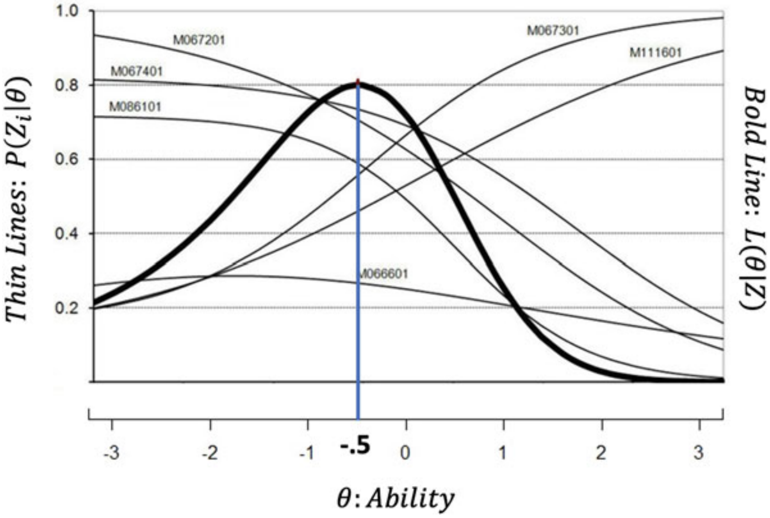


Fig. 2 ICCs and the likelihood (bold) for Student C. The likelihood is calculated by multiplying the student's individual ICCs (Beaton et al., 2011)

shows that Student C's ability would be estimated by maximum likelihood to be about −0.5.

Estimation of ability at the extreme ends of the ability scale is difficult, especially for short tests. Consider Student A in Table 2, who answered all questions incorrectly. His likelihood is shown in Fig. 3 (Beaton et al., 2011). No MLE exists in this case because the likelihood has no maximum. One method for handling estimation for this situation is to assign pre-specified values to examinees who answer no or all questions correctly. This is the method used by the STAAR test in Texas (STAAR, 2004). We will adopt this convention by assigning an ability of −4 to examinees who provide all incorrect responses and an ability of 4 to those who provide all correct responses.

3 Test Information

The test information function (TIF) is defined as the Fisher information of the entire test as a function of ability. One can show that the TIF for the 2PL model, where $p_i(\theta)$ defined in 1, is as defined below:

$$TIF(\theta) = \sum\nolimits_{i=1}^{n} a_i^2 p_i(\theta)(1 - p_i(\theta)). \qquad (2)$$

Two examples of TIFs are presented in Fig. 4a and b. These two curves represent TIFs for tests of ten items that measure ability on a scale that is symmetric around

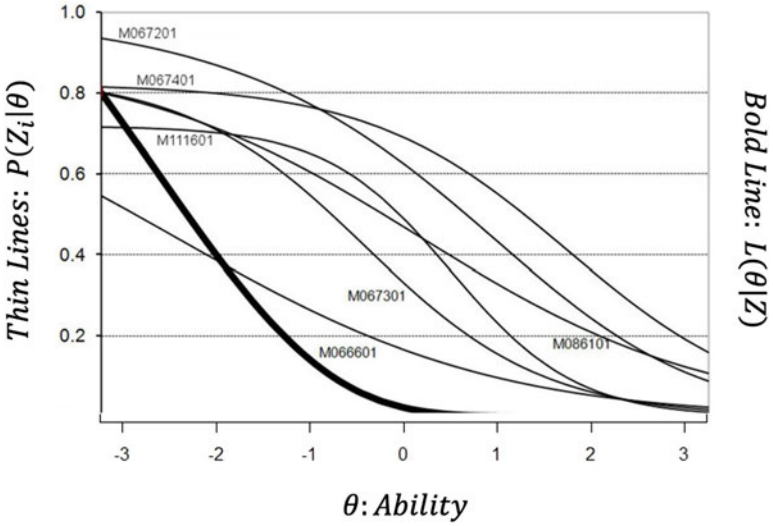


Fig. 3 This plot pictures the ICCs and the likelihood (bold) for Student A. The deficiency of the MLE is exposed in this plot as the student has answered every question incorrectly, and thus the likelihood has no maximum

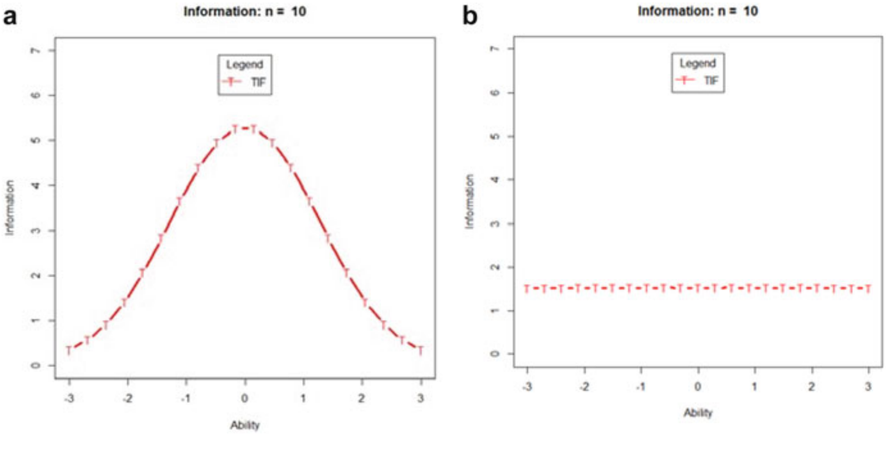


Fig. 4 (**a**) "Peaked" information function; (**b**) Rectangular information function

0, and both will produce some information of examinee ability for those with ability between −3 and +3. However, the tests differ greatly in the shape of their TIFs.

4 Shapes of TIFs

It is common for tests to contain more information about abilities close to the average than at the extremes. The TIF for such a test with ten items[1] is shown in Fig. 4a. It is often desirable that a test maximize information for abilities in the center of the scale, where examinees may be most numerous. This shape is referred to as “peaked." On the other hand, when a population of examinees contains a substantial number at the extremes of the scale, it may be desirable to consider tests with other TIF shapes, such as the “rectangular” one shown in Fig. 4b.

A peaked test information function can be formed through a variety of combinations of items. For instance, a test whose a (discrimination) parameters are similar and whose *b* (difficulty) parameters are grouped near the center will have this shape. On the other hand, a peaked TIF would also result from a test whose *b* parameters are uniformly distributed across the scale and whose a parameters are larger for the items in the center of the range than for those near the tails. Figure 5a displays the discrimination and difficulty parameters of such a test along with its corresponding TIF. Note the increase in item discrimination (a) as the difficulty (b) approaches 0. Figure 5b shows an alternative ten-item test in which the discriminations are nearly constant across the uniformly distributed difficulties which have had a “flattening” effect on the TIF. The tests in Figs. 5a and b will be known as Test 1 and Test 2, respectively, and will be used in examples later in the chapter.

Similar to the peaked TIFs, a rectangular TIF may also be formed through a variety of item parameter combinations. For example, they may have items that have similar a’s and uniformly distributed b’s (Fig. 5c), or they may have more normally distributed b’s, with the items with extreme difficulty having higher a’s

Item	a	b
99	0.79	-2.1
101	0.92	-1.11
14	0.96	-0.61
8	1.02	-0.52
138	1.06	0.15
18	1.68	0.2
161	1.06	0.46
121	1.07	0.49
47	0.97	1.14
155	0.82	2.76

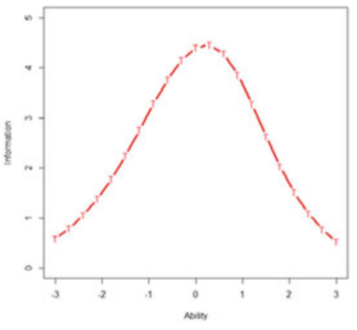


Fig. 5a A ten-item peaked test (Test 1) with uniformly distributed *b* parameters and *a* parameters greater for *b* parameters near 0

[1] These 10 items were real items from the 2004 NAEP Math Exam.

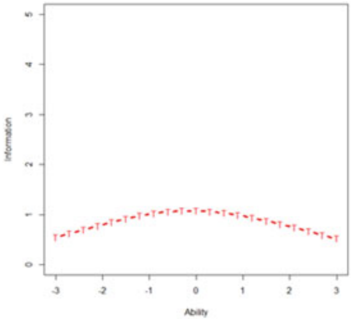
Item	a	b	
82	0.45	-1.73	
147	0.49	-0.72	
17	0.4	-0.42	
32	0.35	-0.259	
117	0.47	-0.25	
118	0.34	0.13	
105	0.41	0.22	
25	0.4	0.87	
136	0.4	1.28	
110	0.46	1.94	

Fig. 5b A ten-item peaked test with uniformly distributed *b* parameters and *a* parameters with less magnitude and nearly uniform across their *b* parameters

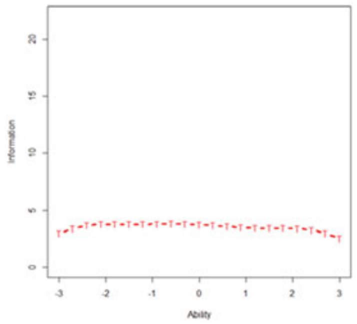
Item	a	b	
85	0.99	-3.1	
92	0.92	-2.11	
15	0.96	-1.61	
76	.902	-0.52	
163	.906	0.15	
144	.968	0.2	
103	.906	0.46	
123	.907	1.09	
87	0.97	2.14	
33	0.87	3.06	

Fig. 5c A 10-item rectangular test (Test 3) with uniformly distributed *a* and *b* parameters

than those near the center. In general, grouping item difficulties and/or increasing item discrimination create peaks in the TIF, while spreading the difficulties and/or decreasing the item discrimination will flatten the TIF. Again, a test with a peaked shaped TIF will be described as a "peaked test," while a test with a flat (rectangular) shaped TIF will be referred to as a "rectangular test."

5 Uses

5.1 Standard Error

An advantage of an IRT model is that its TIF provides an approximate measure of precision for the estimated ability conditional on its value θ:

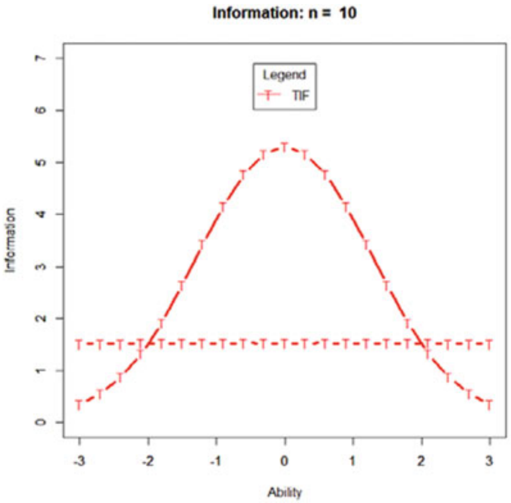


Fig. 6 Peaked and rectangular TIFs superimposed for comparison

$$SE(\theta) = \frac{1}{\sqrt{TIF(\theta)}}.$$

For example, we can see from TIF for the “peaked test" in Fig. 4a that the information provided by the test for an examinee with ability $\theta = 1$ is approximately $I(1) = 4$, yielding an approximate standard error of the ability estimate of $1/\sqrt{4} = 0.5$. However, for a subject of ability $\theta = 2$, $I(2) = 1$ yielding an approximate standard error of $1/\sqrt{1} = 1$. Therefore, this peaked test has less uncertainty for estimated ability of examinees of ability near $\theta = 1$ than for those with ability near $\theta = 2$.

5.2 *Test Construction and Selection*

Another use of the TIF is in item selection and test construction. A test constructor may use the TIFs to choose among tests that measure best for the targeted range of abilities. Figure 6 displays the TIFs from Fig. 4a and b superimposed on one another. If the test constructor is most interested in extremely low or high ability subjects, a rectangular test may be preferred where the information for those examinees is higher. On the other hand, if subjects in the middle of the ability scale make up the target population, the peaked test may be deemed more useful.

6 Small Sample Information of Ability Estimates from IRT Models

As mentioned above, Fisher information measures the asymptotic precision of the maximum likelihood estimator. Therefore, the TIF is a useful tool for standard error estimation and item selection for large tests. An aim of this chapter is to investigate how well it works for that purposes in short tests. Figure 7 shows the TIFs for tests of 10 to 100 items. Each figure shows two curves:

(1) The solid curve is the "actual" test information, defined as the reciprocal of the variance of the MLE and estimated via simulation using the following steps:

Simulation Method for True Information Estimation

(a) An array of quadrature points was created from $\theta = -3$ to $\theta = 3$.
(b) For each quadrature point, a third-party software named MSTSIM5[2] is used to generate 100,000 subjects of that ability as well to simulate each subject's responses to the test of interest.
(c) Each subject's MLE of ability ($\hat{\theta}$) was calculated using MSTSIM5, producing 100,000 estimates of θ for each quadrature point.
(d) The variance of these 100,000 $\hat{\theta}$s ($\widehat{Var}(\hat{\theta})$) was then estimated for each θ in the set of quadrature points.
(e) The true information for each θ in the set of quadrature points was estimated as $\hat{I} = 1/\widehat{Var}(\hat{\theta})$. We will denote this as the actual test information function ($ATIF_{Sim}$).

(2) The dotted curve is the TIF described earlier in (2). This again is the theoretical test information based on an infinitely long test:

As the number of items decrease, the true test information becomes more discrepant from the TIF. In this example, tests of 100 items have information close to what is indicated by the TIF, especially near the center of the curve, but the difference between the two is considerable for smaller tests and for ability levels significantly distant from the center.

However, the discrepancy between the asymptotic and small test size performance is not present for all tests. Figure 8 compares the TIF and the true test information for a rectangular test of ten items. The figure shows that the small sample performance of estimators of ability from this test nearly matches that predicted from asymptotic theory.

To review, we have seen that when a test comprises a large number of items, the TIF is an accurate assessment of its performance. In that case, the asymptotic theory for IRT models is useful and effective for many practical purposes, from assessing

[2] The FORTRAN routine MSTSIM5 (Jodoin, 2003) was used to simulate student responses and calculate the corresponding MLEs for the given IRT models. R was then used to calculate summary statistics (variance, bias, MSE) for these MLEs.

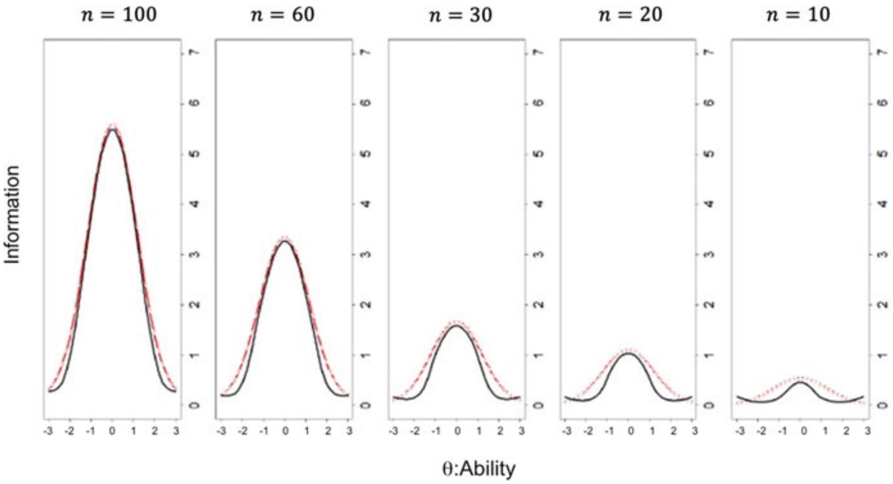
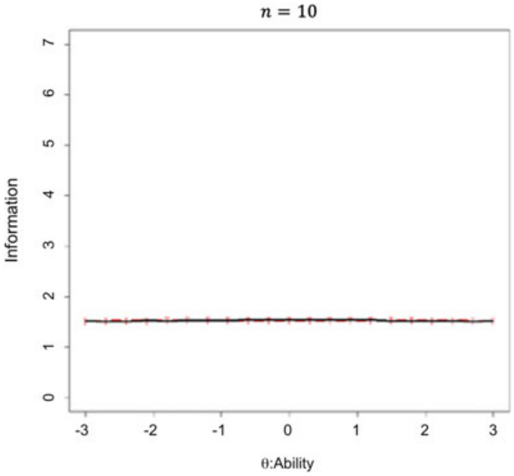


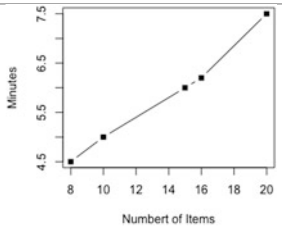
Fig. 7 This figure illustrates how the actual test information (solid black line) increasingly diverges from the theoretical test information (dashed red line) as the test size decreases from $n = 100$ to $n = 10$

Fig. 8 This plot displays the TIF and empirical information for a ten-item test. Compared to Fig. 7, the empirical information is much closer to the TIF which is expected as the TIF is an asymptotic bound of the information



uncertainty in examinee scores to efficient construction of tests. However, there are practical situations when only a few items can be presented to an examinee. One such example is in large-scale assessment, such as the NAEP, where the testing time available is limited. A second example is in multistage testing, where examinees are routed to subsequent stages of varying difficulty based on their performance on earlier stages of the test (Van der Linden & Glas, 2010). Each stage must necessarily consist of a relatively small number of items, after which an ability estimate must be made to facilitate routing. Finally, some tests produce scores on multiples subscales, so that each one may have only a few items. These are the applications in which we

Table 3 Computation times for the simulation method with scatterplot of computation time versus number of items

Simulation method		
Number of Items	Computing Time	
8	4.5 min	
10	5.0 min	
15	6.0 min	
16	6.2 min	
20	7.5 min	

are interested. For "small tests," which we will formally define in a moment, we have seen that the asymptotic theory often overestimates the true test information especially for peaked tests.

We have seen that the method based on simulation can estimate the actual information of the test although it comes with a considerable cost: time. Table 3 shows the computing time of the simulation method to estimate the actual information with 100,000 simulated subjects. All computing was performed on a 4 GB 2.2 GHz Intel i7 processor Apple MacBook Pro for various test sizes and 30 quadrature points. While wait times are subjective, we see that they are at least 4.5 min for an 8-question test and increase linearly with the number of questions at a rate of .24 min per additional item.

7 Exact Method for Information Calculation

Here we provide an alternative to the asymptotically developed TIF and the time-consuming simulation method described above. This method, which we refer to as the exact method, can be broken down into five steps:

1. Generate all possible response patterns given the number of items.
2. Find the unique MLE for each response pattern.
3. For each true ability (discrete number of quadrature points)
   (a) Find the probability for each unique MLE.
   (b) Make a probability distribution given the MLE and corresponding probability from step 3a.

MLE	Probability
$\hat{\theta}_1$	$P(\hat{\theta}_1 \vert \theta)$
$\vdots$	$\vdots$
$\hat{\theta}_{n-1}$	$P(\hat{\theta}_{n-1} \vert \theta)$
$\hat{\theta}_n$	$P(\hat{\theta}_n \vert \theta)$

4. Compute the conditional variance using the equation

$$\sigma^2_{\hat{\theta}} = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta)\right]^2$$

5. Calculate the conditional information as $I(\theta) = \frac{1}{\sigma^2_{\hat{\theta}}}$.

Example Consider a test with the following three items:

Item	a	b
1	1	−2
2	0.5	0
3	0.5	1

Step 1. Generate all possible response patterns given the number of items.

Response pattern	Item 1	Item 2	Item 3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	1	0
6	1	0	1
7	0	1	1
8	1	1	1

Step 2. Find the unique MLE for each response pattern. (From MSTSIM5)

Response pattern	Item 1	Item 2	Item 3	MLE $\hat{\theta}$
1	0	0	0	−4
2	1	0	0	−1.75
3	0	1	0	−2.71
4	0	0	1	−2.71
5	1	1	0	0.92
6	1	0	1	−1.74
7	0	1	1	0.92
8	1	1	1	4

Step 3. For each true ability (discrete number of quadrature points)

(Assume the quadrature points are −3, −2.5, −2, −1.5, −1, −.5, 0, .5, 1, 1.5, 2, 2.5, 3.) We will demonstrate the process for the first quadrature point, $\theta = -3$, and this process would be repeated for each of the remaining 12 quadrature points above.

(a) Find the likelihood (probability) for each unique MLE.

For $\theta = -3$, the probability of response pattern one (missing all three questions) is calculated as

$$\begin{aligned} P(Z|\theta = -3) &= \prod_{i=1}^{3} \left(\frac{1}{1+e^{-1.702a_i(\theta-b_i)}} \right)^{z_i} \left(1 - \frac{1}{1+e^{-1.702a_i(\theta-b_i)}} \right)^{1-z_i} \\ &= \left(\frac{1}{1+e^{-1.702\times 1\times(-3-(-2))}} \right)^{1-0} \times \left(\frac{1}{1+e^{-1.702\times .5\times(-3-(0))}} \right)^{1-0} \\ &\quad \times \left(\frac{1}{1+e^{-1.702\times .5\times(-3-(1))}} \right)^{1-0} \\ &= 0.84580 \times 0.92777 \times 0.96783 = 0.7595. \end{aligned}$$

The probabilities for the remaining 12 quadrature points are found in a similar fashion.

(b) Make a probability distribution given the MLE and likelihood (conditional probability) from step 3a.

For $\theta = -3$,

MLE	$P(\hat{\theta}\|\theta)$
−4	0.7595
−1.75	0.1385
−2.71	0.0591
−2.71	0.0252
0.92	0.0108
−1.74	0.0046
0.92	0.0020
4	0.0004

Step 4. Compute the conditional variance using the equation

$$\sigma^2_{\hat{\theta}} = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta) \right]^2$$

For $\theta = -3$, we have

MLE	$P(\hat{\theta}\|\theta)$	$\hat{\theta}_i^2 P(\hat{\theta}_i\|\theta)$	$\hat{\theta}_i P(\hat{\theta}_i\|\theta)$
−4	0.7595	12.152	−3.038
−1.75	0.1385	0.42415625	−0.242375
−2.71	0.0591	0.43403631	−0.160161
−2.71	0.0252	0.18507132	−0.068292
0.92	0.0108	0.00914112	0.009936
−1.74	0.0046	0.01392696	−0.008004
0.92	0.0020	0.0016928	0.00184
4	0.0004	0.0064	0.0016

$$\sigma^2_{\hat{\theta}} = \sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i^2 P(\hat{\theta}_i|\theta) - \left[\sum_{i=1}^{\text{no. of MLEs}} \hat{\theta}_i P(\hat{\theta}_i|\theta) \right]^2$$

$$= 13.226 - (-3.5034)^2 = 0.952.$$

Step 5. Calculate the conditional information as $I(\hat{\theta}|\theta) = 1/\sigma^2_{\hat{\theta}}$.

For $\theta = -3$, $I(\hat{\theta}|\theta = -3) = 1/0.952 = 1.05$.

Note: in order to find the exact value for a particular ability (i.e., for use in a confidence interval or as a standard error of an estimate), simply follow the steps above and make the quadrature point in step 3 the desired ability.

7.1 Constraint on the Use of the Exact Method

While the exact method yields the exact information/variance for the MLE of ability for any test for which item parameters are known, time is still an important factor. Since the method entails calculating the MLE for every possible response pattern, the number of MLEs to calculate doubles for each item added to the test. This equates to an exponential increase in computation time as the number of items increase. Table 4 shows the computing time of the exact method versus simulation time to estimate the same value with the simulation method. Again, all computing was performed on a 4 GB 2.2 GHz Intel i7 processor Apple MacBook Pro for various test sizes and 30 quadrature points. With a computation time of 2 h, the exact method is practically limited to tests under 20 items. However, since pure simulation is quicker than the exact method beginning at 16 items, we will select the exact method for tests of individual ability with 15 items or fewer.

Table 4 The number of response patterns and computation time for the exact method in calculating the true variance of estimates of individual ability

	Exact method		Simulation method
Number of items	Number of response patterns	Computing time	Computer time
8	$2^8 = 512$	8 s	4.5 min
10	$2^{10} = 1024$	13 s	5 min
15	$2^{15} = 32{,}768$	**3.5 min**	**6 min**
16	$2^{16} = 65{,}536$	**7.33 min**	**6.2 min**
20	$2^{20} = 1{,}048{,}576$	2 h	7.5 min

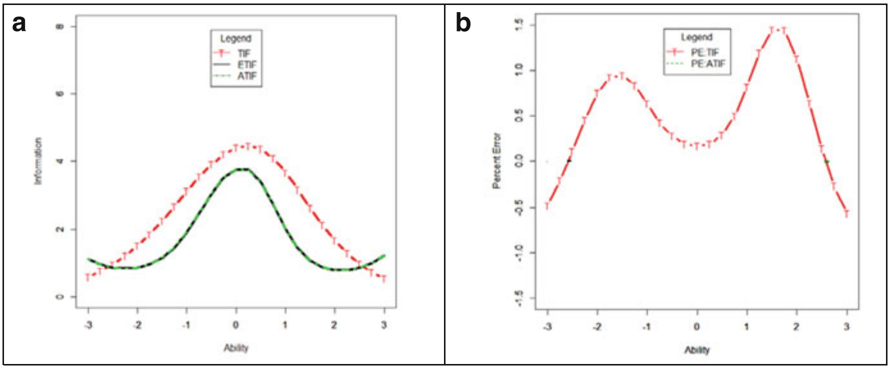


Fig. 9 (**a**) TIF, ATIF, and ETIF for Test 1. The $ATIF_{Exact}$ and $ATIF_{Sim}$ overlap completely; (**b**) PE for the TIF and the ATIF

7.2 Example: Standard Errors

Recall that the square root of the reciprocal of the test information function (TIF) is the asymptotic conditional standard error of the MLE of ability (Hambleton et al., 1991). Some standardized tests, such as the STAAR test in Texas and the CST in California, use square root of the reciprocal of the TIF to report standard errors for their estimates (STAAR, 2004). As we showed above, however, there can be a considerable difference between the TIF and the actual test information. This difference could result in standard errors and confidence intervals that incorrectly represent the variability in the MLE, a particularly troubling problem if the intervals are too narrow.

Figure 9a displays the TIF and the actual information for Test 1 constructed in Fig. 5a. The actual test information is defined as the reciprocal of the true variance of the MLE and was computed by the exact method and is referred to as $ATIF_{Exact}$. For confirmation, the simulated value of the actual information was computed as well, using the simulation method described in the introduction. This function, the $ATIF_{Sim}$, is also shown in Fig. 9a and matches the $ATIF_{Exact}$.

Table 5 The PE with respect to the true SE of Test 1 from the small item bank when the goal is to estimate individual ability

θ	TIF	ATIF	*PE*
−3	0.58	1.12	**−0.48**
−2	1.50	0.86	**0.74**
−1	3.11	1.91	**0.63**
0	4.40	3.77	**0.17**
1	**3.67**	**2.02**	**0.82**
2	1.68	0.79	**1.13**
3	0.54	1.23	**−0.56**

An important note concerns the tails of the $ATIF_{Exact}$ and $ATIF_{Sim}$ in Fig. 9a. As mentioned in the introduction, fixed values are assigned to subjects who obtain perfectly correct and incorrect scores ($\theta = -4$ and $\theta = 4$ were adopted for this study). Therefore, as a subject's ability increases (decreases), a larger percent of them begin to obtain perfectly correct (incorrect) scores and therefore receive an MLE of 4 (−4). This in turn causes a decrease in variance as the true ability approaches 4 (−4), thus resulting in an increase in information. The inflection point of the $ATIF_{Exact}$ and $ATIF_{Sim}$ is the ability level at which subjects begin to obtain perfectly correct (incorrect) scores.

We now examine the difference between the TIF and the ATIF more closely by calculating the percent error (PE) between them:

$$PE = \frac{TIF - ATIF_{Exact}}{ATIF_{Exact}}.$$

Figure 9b displays the PE for the TIF and ATIF (exact and from simulation) in Fig. 9a. Table 5 displays the numerical results. Interestingly, the PE of the TIF is as high as 113%, indicating that the TIF is calculating the information to be 113% higher than it actually is! In a practical setting, the exact method would be used to find the desired standard errors which may then be used in the calculation of confidence intervals.

As an example, consider a fictional subject (Sammy) who was trying to qualify for admission to SMU, where the minimum requirement on the entrance exam is a $\theta = 2.1$.

On a 15-question computer adaptive exam, he received a $\hat{\theta} = 1.0$ and was faced with the decision of whether to retake the exam. Being an asymptotic upper bound on the information, the margin of error using the TIF is smaller than the actual margin of error, thus leading Sammy to believe his true ability is between −0.02 and 2.02 (Table 6); he thus abandons his SMU dream and looks at other schools. However, using the exact method ($ATIF_{Exact}$), we are able to calculate the actual standard error which yields a margin of error of 1.38 (Table 7). Sammy would now be led to believe that his true ability is in the interval (−0.38, 2.38), which contains 2.1 and therefore gives him hope! Although he did not pass the first time, given the actual confidence interval facilitated by the $ATIF_{Exact}$, Sammy receives a more accurate measure of the test's uncertainty and, because he believes passing is now possible, may decide to try the entrance exam a second time.

Table 6 Calculations of the margin of error and 95% confidence limits using the TIF to calculate the SE

Name	Margin of error TIF	95% Confidence interval TIF
Sammy	$1.96 \times \sqrt{3.67} = 1.02$	$1 \pm 1.02 \rightarrow (-0.02, 2.02)$

Table 7 Calculations of the margin of error and 95% confidence limits using the exact method to calculate the exact SE

Name	Margin of error TIF	95% Confidence interval TIF
Sammy	$1.96 \times 1/\sqrt{2.02} = 1.38$	$1 \pm 1.38 \rightarrow (-0.38, 2.38)$

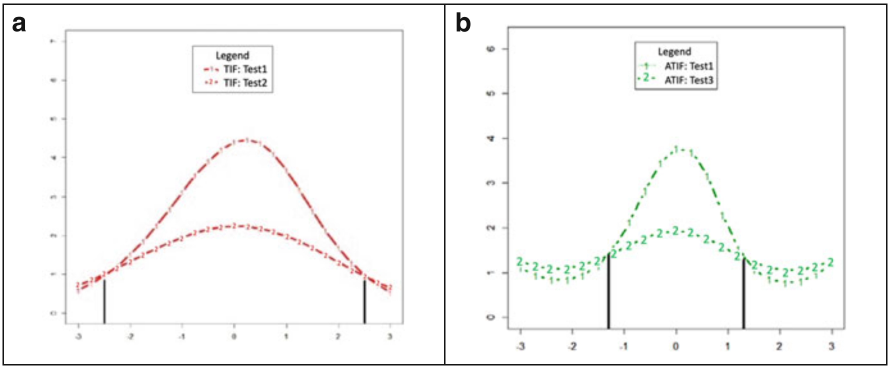


Fig. 10 (**a**) TIFs for Test 1 and Test 2. Test 1 clearly has the higher TIF for the majority of the ability range; (**b**) ATIFs for Test 1 and Test 2. Actual superiority of Test 1 is reduced when the actual information is used

7.3 *Example: Test Construction/Selection*

This example assumes a practitioner would like to compare two tests, both constructed from the NAEP item bank: Test 1 (very peaked from Fig. 5a) and Test 2 (less peaked from Fig. 5b). Figure 10a displays the TIFs from both tests and could be used as a diagnostic tool to decide between them. Assume the practitioner would like to identify students for a remedial math program and has thus been tasked with finding the best test for estimating abilities between −2.5 and −1.5. Judging from the TIFs in 10a, the practitioner would conclude that Test 1 will provide more accurate results because the TIF (the information) is higher over the target range of abilities. We will show, however, that this is not the right conclusion.

We have established that the TIF is an asymptotic target, but this test is only ten items in length. Thus, the practitioner elects to use the exact method to calculate the variance of the estimator and plots the results for both the tests in 10b. The results show that Test 3 is the more accurate test for his target population, as it is superior for $\theta < -1.3$ and $\theta > 1.3$.

8 Conclusion

Calculation of the asymptotic information of estimates of ability in item response theory is useful for tests with a sufficient number of questions. For tests with few items, however, the difference between the theoretical information and the actual information can be substantial. This chapter focused on the practical scenario in which tests have 15 items or fewer. In these cases, the asymptotic estimate can significantly exceed the truth, leading to significant underestimation of the variability of an individual's estimated ability. A relatively quick, exact method of calculating test information can inform test construction and lead to more accurate confidence intervals for individual ability.

References

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). The NAEP primer (NCES 2011–463), U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamental of item response theory*. Sage Publications.

Jodoin, M. (2003). MSTSIM5 [computer software]. University of Massachusetts, School of Education, Amherst, MA.

STAAR2004. (2004). *Technical digest chapter 14: Reliability*. http://www.tea.state.tx.us/student.assessment/techdigest/yr0405/

Van der Linden, W. J., & Glas, C. (2010). *Elements of adaptive testing*. Springer.