Paper SAS364-2014

Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It

Xinming An and Yiu-Fai Yung, SAS Institute Inc.

ABSTRACT

Item response theory (IRT) is concerned with accurate test scoring and development of test items. You design test items to measure various kinds of abilities (such as math ability), traits (such as extroversion), or behavioral characteristics (such as purchasing tendency). Responses to test items can be binary (such as correct or incorrect responses in ability tests) or ordinal (such as degree of agreement on Likert scales). Traditionally, IRT models have been used to analyze these types of data in psychological assessments and educational testing. With the use of IRT models, you can not only improve scoring accuracy but also economize test administration by adaptively using only the discriminative items. These features might explain why in recent years IRT models have become increasingly popular in many other fields, such as medical research, health sciences, quality-of-life research, and even marketing research. This paper describes a variety of IRT models, such as the Rasch model, two-parameter model, and graded response model, and demonstrates their application by using real-data examples. It also shows how to use the IRT procedure, which is new in SAS/STAT® 13.1, to calibrate items, interpret item characteristics, and score respondents. Finally, the paper explains how the application of IRT models can help improve test scoring and develop better tests. You will see the value in applying item response theory, possibly in your own organization!

INTRODUCTION

Item response theory (IRT) was first proposed in the field of psychometrics for the purpose of ability assessment. It is widely used in education to calibrate and evaluate items in tests, questionnaires, and other instruments and to score subjects on their abilities, attitudes, or other latent traits. During the last several decades, educational assessment has used more and more IRT-based techniques to develop tests. Today, all major educational tests, such as the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE), are developed by using item response theory, because the methodology can significantly improve measurement accuracy and reliability while providing potentially significant reductions in assessment time and effort, especially via computerized adaptive testing. In recent years, IRT-based models have also become increasingly popular in health outcomes, quality-of-life research, and clinical research (Hays, Morales, and Reise 2000; Edelen and Reeve 2007; Holman, Glas, and de Haan 2003; Reise and Waller 2009). For simplicity, models that are developed based on item response theory are referred to simply as IRT models throughout the paper.

The paper introduces the basic concepts of IRT models and their applications. The next two sections explain the formulations of the Rasch model and the two-parameter model. Emphases are on the conceptual interpretations of the model parameters. Extensions of the basic IRT models are then described, and some mathematical details of the IRT models are presented. Next, two data examples show the applications of the IRT models by using the IRT procedure. Compared with classical test theory (CTT), item response theory provides several advantages. These advantages are discussed before the paper concludes with a summary.

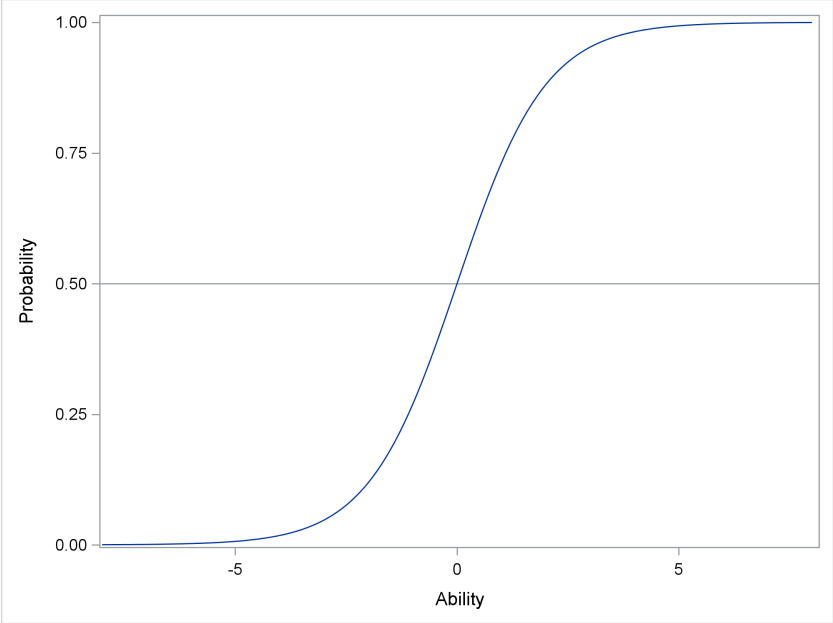
WHAT IS THE RASCH MODEL?

The Rasch model is one of the most widely used IRT models in various IRT applications. Suppose you have J binary items, $X_1, \ldots, X_J$, where 1 indicates a correct response and 0 an incorrect response. In the Rasch model, the probability of a correct response is given by

$$\Pr(x_{ij} = 1) = \frac{e^{\eta_i - \alpha_j}}{1 + e^{\eta_i - \alpha_j}}$$

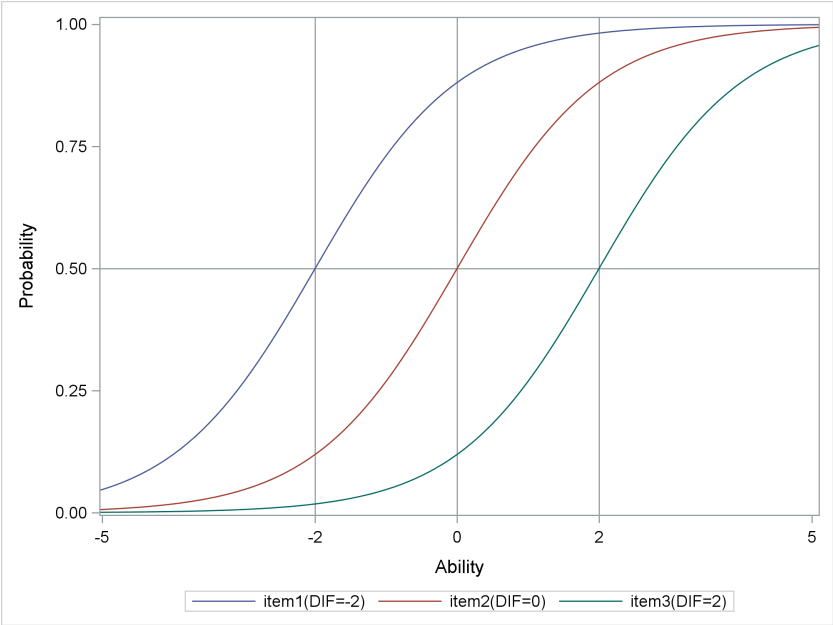
where η_i is the ability (latent trait) of subject i and α_j is the difficulty parameter of item j. The probability of a correct response is determined by the item's difficulty and the subject's ability. This probability can be illustrated by the curve in Figure 1, which is called the item characteristic curve (ICC) in the field of IRT. From this curve you can observe that the probability is a monotonically increasing function of ability. This means that as the subject's ability increases, the probability of a correct response increases; this is what you would expect in practice.

Figure 1 Item Characteristic Curve



As the name suggests, the item difficulty parameter measures the difficulty of answering the item correctly. The preceding equation suggests that the probability of a correct response is 0.5 for any subject whose ability is equal to the value of the difficulty parameter. Figure 2 shows the ICCs for three items, with difficulty parameters of –2, 0, and 2. By comparing these three ICCs, you can see that the location of the ICC is determined by the difficulty parameter. To get a 0.5 probability of a correct response for these three items, the subject must have an ability of –2, 0, and 2, respectively.

Figure 2 Item Characteristic Curves



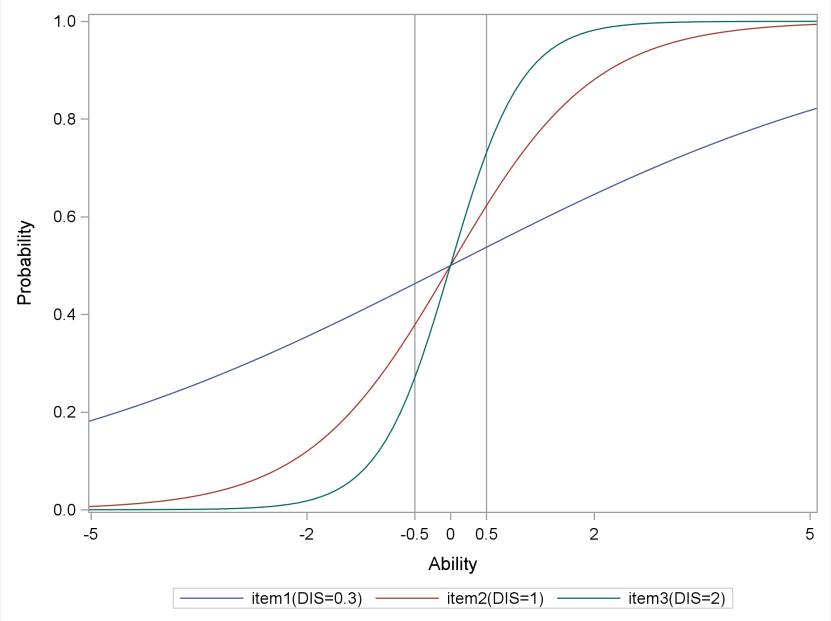
WHAT IS THE TWO-PARAMETER MODEL?

In the Rasch model, all the items are assumed to have the same shape. In practice, however, this assumption might not be reasonable. To avoid this assumption, another parameter called the discrimination (slope) parameter is introduced. The resulting model is called the two-parameter model. In the two-parameter model, the probability of a correct response is given by

$$\Pr(X_{ij} = 1) = \frac{e^{\lambda_j \eta_i - \alpha_j}}{1 + e^{\lambda_j \eta_i - \alpha_j}}$$

where λ_j is the discrimination parameter for item j. The discrimination parameter is a measure of the differential capability of an item. A high discrimination parameter value suggests an item that has a high ability to differentiate subjects. In practice, a high discrimination parameter value means that the probability of a correct response increases more rapidly as the ability (latent trait) increases. Item characteristic curves of three items, **item1**, **item2**, and **item3**, with different discrimination parameter values are shown in Figure 3.

Figure 3 Item Characteristic Curves



The difficulty parameter values for these three items are all 0. The discrimination parameter values are 0.3, 1, and 2, respectively. In Figure 3, you can observe that as the discrimination parameter value increases, the ICC becomes more steep around 0. As the ability value changes from –0.5 to 0.5, the probability of a correct response changes from 0.3 to 0.7 for **item3**, which is much larger than **item1**. For that reason, **item3** can differentiate subjects whose ability value is around 0 more efficiently than **item1** can.

EXTENSIONS OF THE BASIC IRT MODELS

Early IRT models, such as the Rasch model and the two-parameter model, concentrate mainly on analyzing dichotomous responses that have a single latent trait. The preceding sections describe the characteristics of these two models. Various extensions of these basic IRT models have been developed for more flexible modeling in different situations. The following list presents some extended (or generalized) IRT models and their capabilities:

- graded response models (GRM), which analyze ordinal responses and rating scales
- three- and four-parameter models, which analyze test items that have guessing and ceiling parameters in the response curves

- multidimensional IRT models, which analyze test items that can be explained by more than one latent trait or factor
- multiple-group IRT models, which analyze test items in independent groups to study differential item functioning or invariance
- confirmatory IRT models, which analyze test items that have hypothesized relationships with the latent factors

These generalizations or extensions of IRT models are not mutually exclusive. They can be combined to address the complexity of the data and to test the substantive theory in practical applications. Although the IRT procedure handles most of these complex models, it is beyond the scope of this paper to describe all these models in detail. For general references about various IRT models, see De Ayala (2009) and Embretson and Reise (2000). The current paper focuses on the basic unidimensional IRT models that are used in the majority of applications.

SOME MATHEMATICAL DETAILS OF IRT MODELS

This section provides mathematical details of the multidimensional graded response model for ordinal items. This model subsumes most basic IRT models, such as the Rasch model and the two-parameter model, as special cases. Mathematically inclined readers might find this section informative, whereas others might prefer to skip it if their primary goal is practical applications.

A d-dimensional IRT model that has J ordinal responses can be expressed by the equations

$$y_{ij} = \boldsymbol{\lambda}_j \boldsymbol{\eta}_i + \epsilon_{ij}$$

$$p_{ijk} = \Pr(u_{ij} = k) = \Pr(\alpha_{(j,k-1)} < y_{ij} < \alpha_{(j,k)}), \quad k = 1, \ldots, K$$

where u_{ij} is the observed ordinal response from subject i for item j; y_{ij} is a continuous latent response that underlies u_{ij}; $\boldsymbol{\alpha}_j = (\alpha_{(j,0)} = -\infty, \alpha_{(j,1)}, \ldots, \alpha_{(j,K-1)}, \alpha_{(j,K)} = \infty)$ is a vector of threshold parameters for item j; $\boldsymbol{\lambda}_j$ is a vector of slope (or discrimination) parameters for item j; $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{id})$ is a vector of latent factors for subject i, $\boldsymbol{\eta}_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ is a vector of unique factors for subject i. All the unique factors in $\boldsymbol{\epsilon}_i$ are independent of one another, suggesting that $y_{ij}, j = 1, \ldots, J$, are independent conditional on the latent factor $\boldsymbol{\eta}_i$. (This is the so-called local independence assumption.) Finally, $\boldsymbol{\eta}_i$ and $\boldsymbol{\epsilon}_i$ are also independent.

Based on the preceding model specification,

$$p_{ijk} = \int_{\alpha_{(j,k-1)}}^{\alpha_{(j,k)}} p(y; \lambda_j \eta_i, 1) dy = \int_{\alpha_{(j,k-1)} - \lambda_j \eta_i}^{\alpha_{(j,k)} - \lambda_j \eta_i} p(y; 0, 1) dy$$

where p is determined by the link function. It is the density function of the standard normal distribution if the probit link is used, or the density function of the logistic distribution if the logistic link is used.

The model that is specified in the preceding equation is called the multidimensional graded response model. When the responses are binary and there is only one latent factor, this model reduces to the two-parameter model, which can be expressed as

$$y_{ij} = \lambda_j \eta_i + \epsilon_{ij}$$

$$p_{ij} = \Pr(u_{ij} = 1) = \Pr(y_{ij} > \alpha_j)$$

A different parameterization for the two-parameter model is

$$y_{ij} = a_j(\eta_i - b_j) + \epsilon_{ij}$$

$$p_{ij} = \Pr(u_{ij} = 1) = \Pr(y_{ij} > 0)$$

where b_j is interpreted as item difficulty and a_j is called the *discrimination parameter*. The preceding two parameterizations are mathematically equivalent. For binary response items, you can transfer the threshold parameter into the difficulty parameter by $b_j = \frac{\alpha_j}{\lambda_j}$. The IRT procedure uses the first parameterization.

The two-parameter model reduces to a one-parameter model when slope parameters for all the items are constrained to be equal. In the case where the logistic link is used, the one- and two-parameter models are often abbreviated as 1PL and 2PL, respectively. When all the slope parameters are set to 1 and the factor variance is set to a free parameter, you obtain the Rasch model.

You can obtain three- and four-parameter models by introducing the guessing and ceiling parameters. Let g_j and c_j denote the item-specific guessing and ceiling parameters, respectively. Then the four-parameter model can be expressed as

$$p_{ij} = \Pr(u_{ij} = 1) = g_j + (c_j - g_j)\Pr(y_{ij} > 0)$$

This model reduces to the three-parameter model when $c_j = 1$.

EXAMPLE 1: LAW SCHOOL ADMISSION TEST

The data set in this example comes from the Law School Admission Test (LSAT). It includes the responses of 1,000 subjects to five binary items. The following DATA step creates the data set IrtLsat:

```
data IrtLsat;
   input item1-item5 @@;
   datalines;
0 0 0 0 0

   ... more lines ...

1 1 1 1 1
;
```

The following statements fit the IRT model by using all the default settings. The PLOTS= option is used to request the scree plot and the item characteristic curves, with the arguments SCREE and ICC.

```
ods graphics on;
proc irt data=IrtLsat plots=(scree icc);
   var item1-item5;
run;
```

The unidimensional assumption suggests that the correlation among these items can be explained by a single latent factor. You can check this assumption by examining the eigenvalues and the magnitude of the item slope (discrimination) parameters. A small slope parameter value (such as < 0.5) often suggests that the corresponding item is not a good indicator of the latent construct. Figure 4 and Figure 5 show the eigenvalue table and the scree plot, respectively. You can see that the first eigenvalue is much greater than the others, suggesting that a unidimensional model is reasonable for this example.

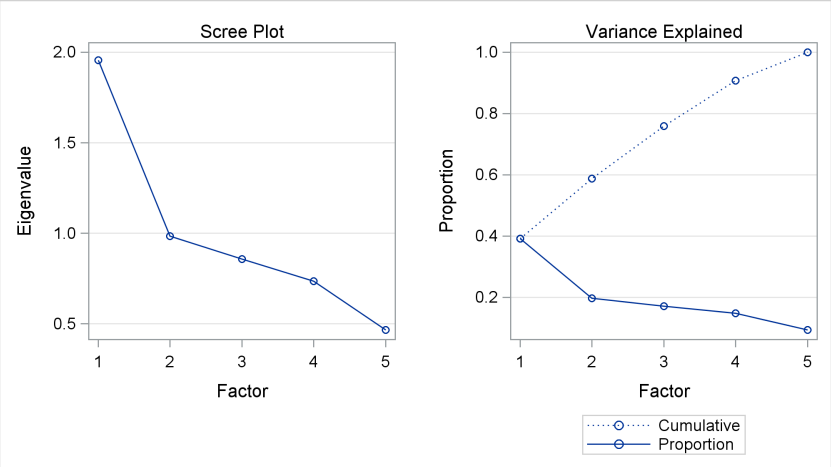
Figure 4 Eigenvalues of Polychoric Correlations

The IRT Procedure

Eigenvalues of the Polychoric Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.95547526	0.97064793	0.3911	0.3911
2	0.98482733	0.12784702	0.1970	0.5881
3	0.85698031	0.12009870	0.1714	0.7595
4	0.73688161	0.27104612	0.1474	0.9068
5	0.46583549		0.0932	1.0000

Figure 5 Scree Plots



Parameter estimates for this example are shown in Figure 6. Under the parameterization used by PROC IRT, the slope parameter is the same as the discrimination parameter. As a result, these parameters are used interchangeably throughout this paper. The threshold parameter has the same interpretation as the difficulty parameter. For this example, the threshold parameter estimates range from –2.59 to –0.19; **item1** is the easiest item, and **item3** is the most difficult item. The fact that all the threshold parameter estimates are less than 0 suggests that all the items in this example are relatively easy and therefore are most useful in discriminating subjects who have lower abilities. As mentioned in the preceding section, the threshold parameter can be transformed into the difficulty parameter. For each ICC plot shown in Figure 7, the vertical reference line indicates the difficulty of each item. The difficulty parameter value is shown at the top of each plot beside the vertical reference line.

The slope parameter values for this example range from 0.45 to 1.36. By comparing the ICCs in Figure 7, you can observe how the value of the slope parameter affects the shape of the ICC. Among these five items, the ICC for **item1** is the steepest and the ICC for **item3** is the flattest.

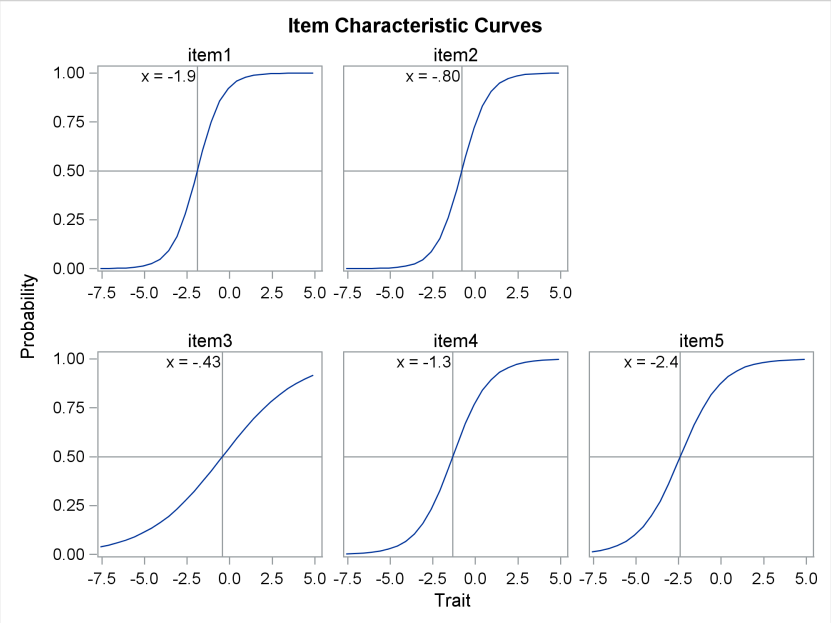
Figure 6 Parameter Estimates

The IRT Procedure

Item Parameter Estimates

Item	Parameter	Estimate	Standard Error	Pr > |t|
item1	**Threshold**	-2.59087	0.22115	<.0001
	Slope	1.36225	0.25067	<.0001
item2	**Threshold**	-1.05859	0.12506	<.0001
	Slope	1.32388	0.27282	<.0001
item3	**Threshold**	-0.19313	0.06667	0.0019
	Slope	0.44845	0.11478	<.0001
item4	**Threshold**	-1.26496	0.10733	<.0001
	Slope	0.95289	0.18798	<.0001
item5	**Threshold**	-1.98140	0.12915	<.0001
	Slope	0.82665	0.17174	<.0001

Figure 7 Item Characteristic Curves



EXAMPLE 2: QUALITY OF LIFE SURVEY

The data set in this example comes from the 1978 Quality of American Life Survey. The survey was administered to a sample of all US residents aged 18 years and older in 1978. Subjects were asked to rate their satisfaction with many different aspects of their lives. This example includes 14 items. Some of the items are as follows:

- satisfaction with community
- satisfaction with neighbors
- satisfaction with amount of education received
- satisfaction with health
- satisfaction with job
- satisfaction with income

This example uses 1,000 random samples from the original data set. The following DATA step creates the data set IrtSat:

```
data IrtSat;
   input item1-item14 @@;
   datalines;
 1   1   2   1   1    2    2    2    .    2    2    2    2    2

   ... more lines ...

 1   1   1   1   2    2    2    2    .    1    1    1    1    3
;
```

For illustration purposes, these items are reorganized into different numbers of categories. The number of categories ranges from 2 to 7. By default, the IRT procedure uses the graded response model (GRM) with the logistic link for all the items. For binary response, the GRM is equivalent to the two-parameter model. In

PROC IRT, you can specify different types of response models for different items by using the RESFUNC= option in the MODEL statement.

For this example, because all the items are designed to measure subjects' satisfaction with different aspects of their lives, it is reasonable to start with a unidimensional IRT model. The following statements fit such a model by using all the default options:

```
ods graphics on;
proc irt data=IrtSat plots=(iic tic) ;
   var item1-item14;
run;
```

Figure 8 shows the eigenvalue table for this example. You can see that the first eigenvalue is much greater than the others, suggesting that a unidimensional model is reasonable for this example.

Figure 8 Eigenvalues of Polychoric Correlations

The IRT Procedure

Eigenvalues of the Polychoric Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.57173396	4.19614614	0.3980	0.3980
2	1.37558781	0.29273244	0.0983	0.4962
3	1.08285537	0.12600033	0.0773	0.5736
4	0.95685504	0.09108909	0.0683	0.6419
5	0.86576595	0.09758221	0.0618	0.7038
6	0.76818374	0.12571683	0.0549	0.7586
7	0.64246691	0.06108305	0.0459	0.8045
8	0.58138386	0.04214553	0.0415	0.8461
9	0.53923833	0.10092835	0.0385	0.8846
10	0.43830998	0.07346977	0.0313	0.9159
11	0.36484021	0.04667935	0.0261	0.9419
12	0.31816085	0.03905135	0.0227	0.9647
13	0.27910950	0.06360101	0.0199	0.9846
14	0.21550849		0.0154	1.0000

In the context of IRT, the amount of information that each item or test provides is not evenly distributed across the entire continuum of latent constructs. The value of the slope parameter represents the amount of information provided by the item. For this example, parameter estimates and item information curves are shown in Figure 9 and Figure 10, respectively. By examining the parameter estimates and the item information curves, you can see that items that have high slope values are more informative than items that have low slope values. For example, because the slope value of **item1** is much smaller than the slope value of **item9**, the information curve is flatter for **item1** than for **item9**.

For individual items, most of the information concentrates around the area defined by the threshold parameters. The binary response item provides most of the information around the threshold. For ordinal items, most of the information falls in the range defined by the lowest and the highest threshold parameters. By comparing the information curves for **item7** and **item9**, you can also see that in cases where response items have the same slope value, the ordinal item is more informative than the binary item.

Item selection is an important process for test (questionnaire) development. It serves two purposes: to ensure that all the items included in the test are sufficiently unidimensional, and to maximize the test information across the interested continuum of latent constructs. During the item selection process, ideally you want to select high-differential items whose threshold parameters cover the interested latent construct continuum. However, in practice you often encounter the situation in which these high-differential items cannot provide enough information for the entire continuum, especially when these items are binary. In this situation, you might need to select some lower-differential items that can add information to the area that is not covered by these high-differential items.

Figure 9 Parameter Estimates

The IRT Procedure

Item Parameter Estimates

Item	Parameter	Estimate	Standard Error	Pr > |t|
item1	Threshold 1	-0.95503	0.07367	<.0001
	Threshold 2	1.47924	0.08372	<.0001
	Slope	0.45394	0.07042	<.0001
item2	Threshold 1	-0.65531	0.08328	<.0001
	Threshold 2	0.56898	0.08242	<.0001
	Slope	1.20647	0.09708	<.0001
item3	Threshold	0.00170	0.07114	0.4905
	Slope	0.72428	0.08511	<.0001
item4	Threshold	-0.44155	0.08367	<.0001
	Slope	1.23650	0.11147	<.0001
item5	Threshold 1	-0.73837	0.08045	<.0001
	Threshold 2	0.64113	0.07930	<.0001
	Slope	1.03810	0.08755	<.0001
item6	Threshold 1	-0.42252	0.07133	<.0001
	Threshold 2	0.85163	0.07544	<.0001
	Slope	0.70786	0.07605	<.0001
item7	Threshold 1	-1.47166	0.11278	<.0001

The IRT Procedure

Item Parameter Estimates

Item	Parameter	Estimate	Standard Error	Pr > |t|
	Threshold 2	0.48414	0.10006	<.0001
	Threshold 3	1.68936	0.11665	<.0001
	Slope	1.89361	0.12255	<.0001
item8	Threshold 1	-1.04483	0.09261	<.0001
	Threshold 2	0.87777	0.09072	<.0001
	Threshold 3	1.94715	0.11031	<.0001
	Slope	1.41242	0.09836	<.0001
item9	Threshold	0.33440	0.11941	0.0026
	Slope	1.85067	0.20122	<.0001
item10	Threshold 1	-0.54424	0.09447	<.0001
	Threshold 2	1.15923	0.10220	<.0001
	Slope	1.66943	0.12080	<.0001
item11	Threshold 1	-1.69919	0.11397	<.0001
	Threshold 2	-0.03551	0.09403	0.3529
	Threshold 3	1.09760	0.10224	<.0001
	Threshold 4	2.27404	0.12716	<.0001
	Slope	1.66135	0.10896	<.0001

Figure 9 *continued*

The IRT Procedure

Item Parameter Estimates

Item	Parameter	Estimate	Standard Error	Pr > |t|
item12	Threshold 1	-2.22687	0.11781	<.0001
	Threshold 2	-0.95057	0.08845	<.0001
	Threshold 3	-0.10165	0.08221	0.1082
	Threshold 4	0.73278	0.08572	<.0001
	Threshold 5	1.48510	0.09659	<.0001
	Threshold 6	2.20842	0.11444	<.0001
	Slope	1.18861	0.08590	<.0001
item13	Threshold 1	-1.99808	0.14409	<.0001
	Threshold 2	0.81878	0.12153	<.0001
	Threshold 3	2.72939	0.16482	<.0001
	Slope	2.47742	0.15809	<.0001
item14	Threshold 1	-1.90189	0.10988	<.0001
	Threshold 2	0.51378	0.08654	<.0001
	Threshold 3	1.91870	0.10854	<.0001
	Slope	1.39119	0.09549	<.0001

Figure 10 Item Information Curves

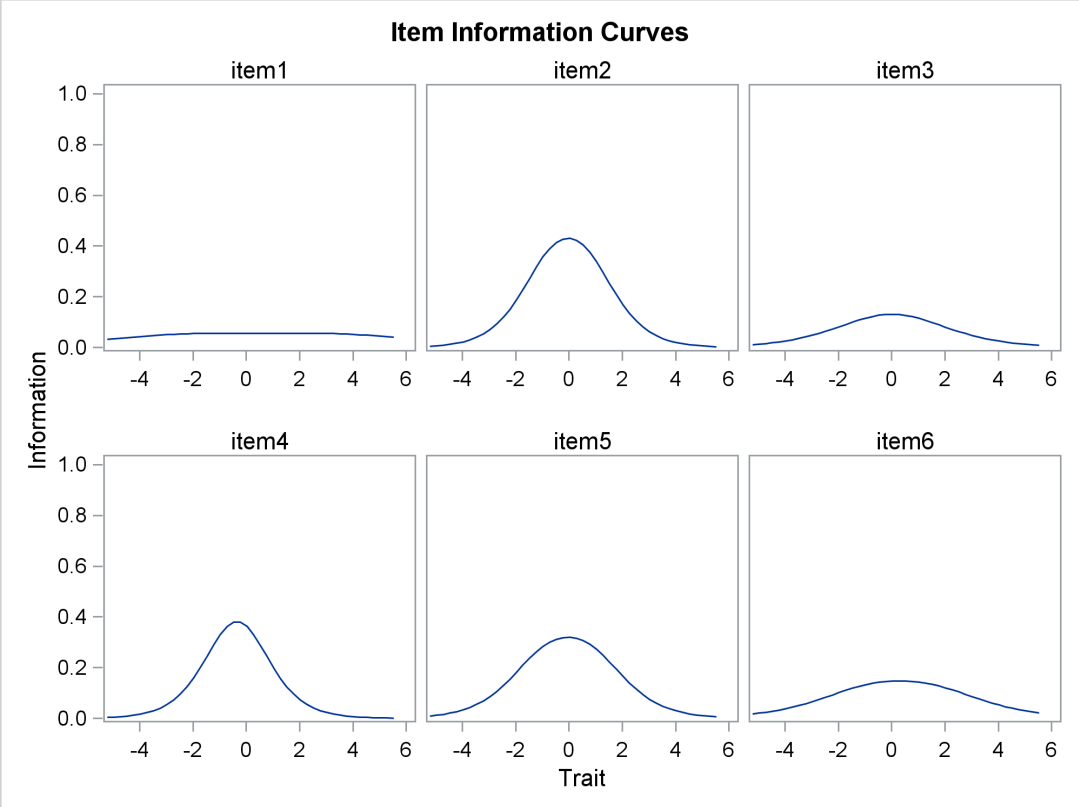
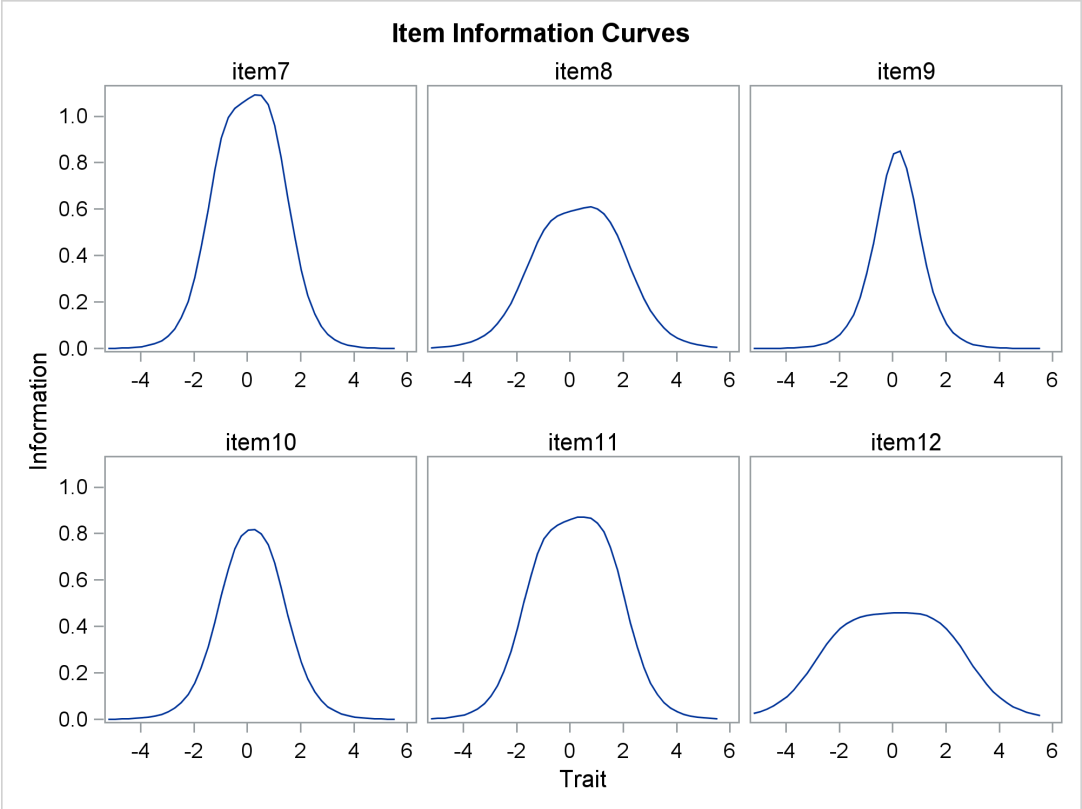
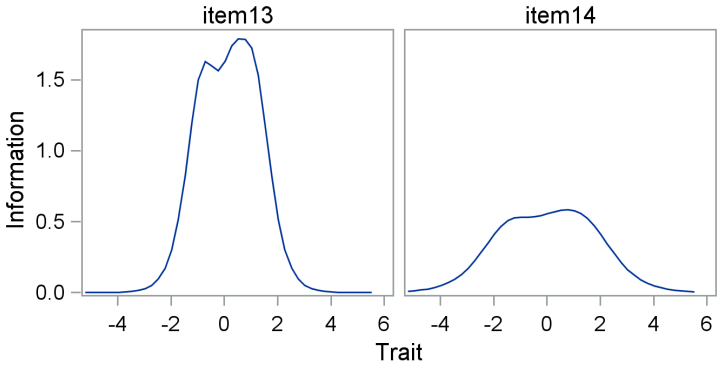


Figure 10 *continued*



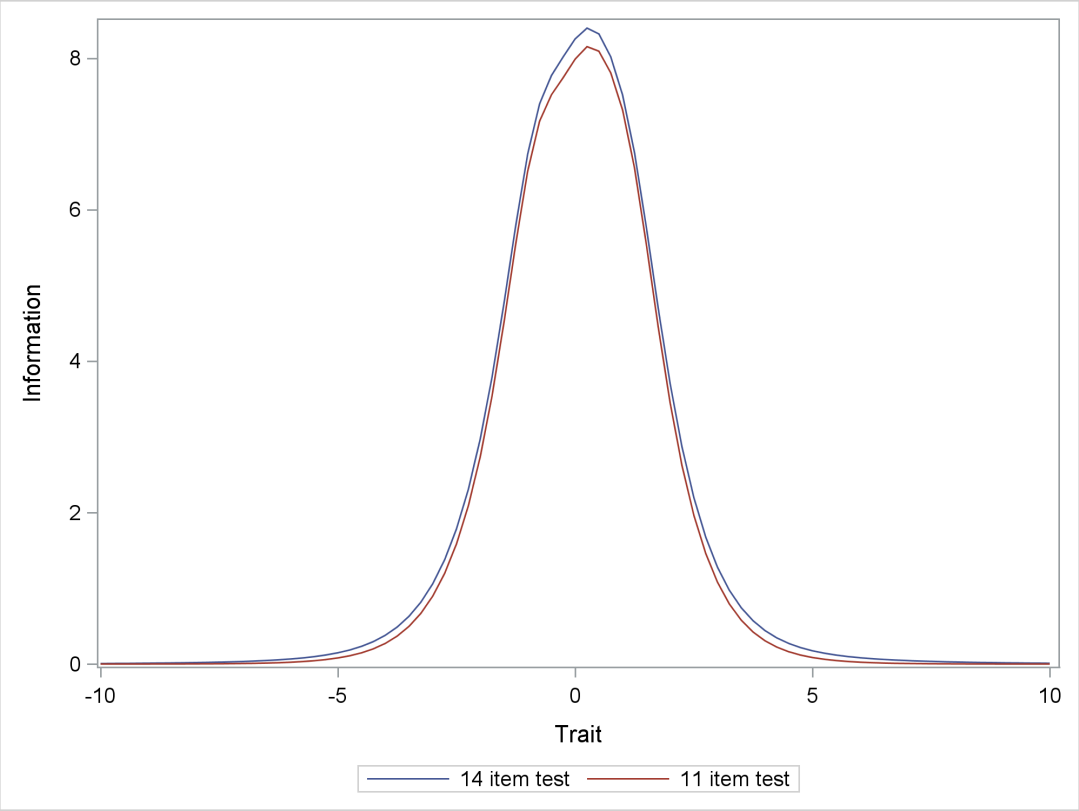


For this example, the slope parameters range from 0.45 to 2.47, and the threshold parameters range from –2.23 to 2.47. Among these 14 items, three of them, **item1**, **item3**, and **item6**, have slope values below 1, and the slope value for **item1** is below 0.5. The item information curves suggest that these three items, especially **item1**, provide much less information than the other items. As a result, you might consider dropping these three items. Figure 11 shows the test information curves for the original test that has 14 items and the shorter test that excludes **item1**, **item3**, and **item6**. The two information curves are almost identical, suggesting that the shorter test provides almost the same amount of information as the longer test.

After item calibration and item selection, another important task is to score subjects based on their responses. In IRT, there are three widely used scoring methods: maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP). You can specify them by using the SCOREMETHOD= option in the PROC IRT statement. The following example code scores the subject based on the shorter test by using the MAP method:

```
proc irt data=IrtSat scoremethod=map;
   var item2 item4 item5 item7-item14;
run;
```

Figure 11 Test Information Curves



ITEM RESPONSE THEORY—DOES IT OFFER MORE THAN CTT CAN PROVIDE?

Classic test theory (CTT) has been the basis for developing psychological scales and test scoring for many decades. You might wonder why you would need IRT models, which serve similar purposes. This section points out some conceptual and practical advantages of IRT models over CTT models in regard to test or scale construction and development. The main purpose is to point out that IRT models have unique features that complement CTT-based measures rather than to thoroughly compare the two approaches.

First, a limitation of CTT is that the item and person characteristics, such as item difficulty parameters and person scores, are not discernible. Depending on the subpopulation in question, item characteristics might change. If a high-ability subpopulation is considered, all test items would appear to be easy. But when a

low-ability subpopulation is considered, the same set of items would be difficult. This limitation makes it difficult to assess individuals' abilities by using different test forms. However, in IRT, the item characteristics and the personal abilities are formulated by distinctive parameters. After the items are calibrated for a population, the scores for subjects from that population can be compared directly even if they answer different subsets of the items. Some researchers refer to this as the invariant property of IRT models (for example, see Hambleton, Swaminathan, and Rogers 1991).

Second, the definition of reliability in CTT is based on parallel tests, which are difficult to achieve in practice. The precision of measurement is the same for all scores for a particular sample. In CTT, longer tests are usually more reliable than shorter tests. However, reliability in IRT is defined as a function that is conditional on the scores of the measured latent construct. Precision of measurement differs across the latent construct continuum and can be generalized to the whole target population. In IRT, measurement precision is often depicted by the information curves. These curves can be treated as a function of the latent factor conditional on the item parameters. They can be calculated for an individual item (item information curve) or for the whole test (test information curve). The test information curve can be used to evaluate the performance of the test. During test development, you want to make sure that the selected items can provide adequate precision across the interested range of the latent construct continuum.

Third, missing values in CTT are difficult to handle during both test development and subject scoring. Subjects who have one or more missing responses cannot be scored unless these missing values are imputed. In contrast, the estimation framework of IRT models makes it straightforward to analyze items that have random missing data. IRT can still calibrate items and score subjects by using all the available information based on the likelihood; the likelihood-based methods are implemented in the IRT procedure.

MAIN FEATURES OF THE IRT PROCEDURE

The IRT procedure enables you to estimate various IRT models. The following list summarizes some of the main features of PROC IRT:

- fits the following classes of models: Rasch model; one-, two-, three-, and four-parameter models; and graded response models
- supports logistic and probit links
- calibrates items that can have different response models
- performs multidimensional exploratory and confirmatory analysis
- performs multiple-group analysis, with fixed values and equality constraints within and between groups
- estimates factor scores by using the maximum likelihood (ML), maximum a posteriori (MAP), or expected a posteriori (EAP) method
- supports the quasi-Newton (QN) and expectation-maximization (EM) algorithms for optimization

SUMMARY

This paper provides a brief introduction to item response theory (IRT) and the related models. The Rasch and two-parameter models are the two models most frequently used in applications. Basic concepts and interpretations of these basic models are described. Examples illustrate the use of the IRT procedure, which is new in SAS/STAT 13.1, for fitting IRT models and selecting useful items. IRT provides a modeling framework that you can use to study item characteristics and person scores unambiguously from the data. Some advantages of IRT over classic testing theory (CTT) are discussed. The data examples in this paper illustrate only some of the functionality that PROC IRT provides. Actual test developments might require more extensive analysis, including measures based on CTT and inputs from content experts. For detailed discussions of test development, see DeVellis (2011) and Edelen and Reeve (2007).

REFERENCES

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

DeVellis, R. F. (2011), *Scale Development: Theory and Applications*, 3rd Edition, Thousand Oaks, CA: Sage Publications.

Edelen, M. O. and Reeve, B. B. (2007), "Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement," *Quality of Life Research*, 16, 5–18.

Embretson, S. E. and Reise, S. P. (2000), *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991), *Fundamentals of Item Response Theory*, Newbury Park, CA: Sage Publications.

Hays, R. D., Morales, L. S., and Reise, S. P. (2000), "Item Response Theory and Health Outcomes Measurement in the Twenty-First Century," *Medical Care*, 38, Suppl. 9, 1128–1142.

Holman, R., Glas, C. A. W., and de Haan, R. J. (2003), "Power Analysis in Randomized Clinical Trials Based on Item Response Theory," *Controlled Clinical Trials*, 24, 390–410.

Reise, S. P. and Waller, N. G. (2009), "Item Response Theory and Clinical Measurement," *Annual Review of Clinical Psychology*, 5, 27–48.

ACKNOWLEDGMENTS

The authors are grateful to Bob Rodriguez and Ed Huddleston of the Advanced Analytics Division at SAS Institute Inc. for their valuable assistance in the preparation this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Xinming An	Yiu-Fai Yung
SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513
Xinming.An@sas.com	Yiu-Fai.Yung@sas.com