Course: Test Development and Evaluation (6462)
Semester: Autumn, 2021

ASSIGNMENT No. 2

Q.1 What are the characteristics of classroom projects?

There are many kinds of projects implemented in classrooms. Effective projects balance the level of learner control with teacher-planned structure that guides and focuses learner work. The characteristics below help define effective project-based learning experiences.

Leareners are at the centre of the learning process.

Well designed project-based experiences engage learners in open-ended, authentic tasks. Compelling project tasks empower learners to make decisions and apply their interests and passions to culminating products and performances. Learners learn through inquiry and have some control over decisions about how they complete project tasks. The teacher takes on the role of a facilitator or coach. Learners often work in collaborative groups, assuming roles that make best use of their individual talents.

Projects focus on learning that is aligned with assessment standards.

Good projects are developed around core curricular concepts that address national assessment standards. The project has clear goals that align with assessment standards and focus on what learners should be able to understand and do as a result of their learning. With a focus on assessment standards, the teacher defines appropriate demonstrations of learning in an assessment plan and organizes learning activities and teaching interventions. Project work culminates in learner products and performance tasks such as persuasive presentations and informational newsletters that demonstrate understanding and achieve assessment standards .

Projects are driven by Curriculum-Aligned Questions.

Questions keep projects focused on important learning. Learners are challenged to dig deeper with subject-specific content questions that structure their research. There are two types of Curriculum-Aligned Questions: Focus and Content Questions. Focus Questions are tied directly to the project and support investigation in a focused way. Focus Questions help demonstrate how well learners understand the core concepts of the project. Content Questions are more fact-based and ask the more basic questions that are necessary to engage with the more challenging Focus Question.

Projects involve on-going and multiple types of assessment.

Clear expectations are defined at the beginning of a project and are revisited with multiple checks for understanding using varied assessment methods. Learners have models and guidelines for high quality work and know what is expected of them from the beginning of the project. Opportunities for reflection, feedback, and adjustment are embedded in the project.

The project has real-world connections.

Projects are relevant to learners ' lives and may involve community or outside experts who provide a context for learning. Learners may present their learning to an authentic audience, connect with community resources, tap into experts in the field of study, or communicate through technology.

Learners demonstrate knowledge through a product or performance.

Projects typically culminate with learners demonstrating their learning through presentations, written documents, constructed displays, proposals, or even simulated events such as a mock trial. These final products allow for learner expression and ownership of learning.

ICT supports and enhances learning.

Learners have access to different types of ICT, which are used to support the development of thinking skills, content expertise, and creation of final products. With the help of ICT, learners have more control over final results and an opportunity to personalize products. Learners can reach beyond the walls of the classroom by collaborating with distant classes through email and self-made websites, or presenting their learning through multimedia.

Thinking skills are integral to project work.

Project work supports the development of both metacognitive and cognitive thinking skills such as collaboration, self-monitoring, analysis of data, and evaluation of information. Throughout the project, Curriculum-Aligned Questions challenge learners to think and make connections to concepts that matter in the real world.

Teaching and learning strategies are varied and support multiple learning styles. Teaching and learning strategies create a richer learning environment and promote higher-order thinking. A range of strategies ensures that the curricular material is accessible to all learners and provides opportunities for every learner to succeed. Teaching and learning may include the use of different cooperative grouping strategies, graphic organizers, and teacher and peer feedback.

Q.2 Critically analyze the oral questioning? Also give guidelines to develop these.

1. When presenting a lesson, you have many techniques and aids at your disposal. One aid that can be used to stimulate learning and effectively applied to satisfy all seven learning factors is oral questioning.
2. The actual technique of questioning is a difficult one and is normally one of the most neglected areas of instruction. Good oral questioning requires the ability to think quickly and easily while facing a class or individual student, to shift and change as thought progresses and to phrase questions in clear and simple terms. You must always be mindful of the technique to follow when handling student questions and answers.

Purposes of Oral Questions

1. First, questions can be used to PROMOTE MENTAL ACTIVITY. You can state a fact and provide visual or verbal support to back it up but the surest way for students to remember is working it out for themselves. Whenever you can use an oral question to make your students think and reason out the fact, you should take advantage of the situation. Example: As students work towards an objective it is often necessary for them to recall pertinent data or knowledge learned previously. A well-worded oral question could provide the required information, thus promoting mental activity.

2. A second purpose of oral questions is to AROUSE AND MAINTAIN STUDENT INTEREST. Merely making a statement will often result in a "so what" attitude, but asking questions makes students feel they are participating and contributing to the lesson, thereby arousing interest. You can maintain this interest throughout the lesson by the continuous development of facts and ideas. Remember? Telling is NOT teaching.
3. Another purpose of oral questions is to GUIDE THOUGHT. By using questions you can lead students to think through to a logical solution. Questions can direct students' thinking through a definite sequence or to particular objectives. During discussions you can use questions to guide your student's thoughts back to the objective if they seem to be far afield. An experienced instructor can guide students through an entire lesson by asking the right questions at the right time.
4. A final purpose of oral questions is to EVALUATE LEARNING for the benefit of both instructor and student. Oral questions may be used after each stage of a lesson to ensure students are following before you proceed to the next stage. At the end of the lesson, they confirm that students have attained the objectives for that particular lesson.

A drawback of using oral questions to evaluate learning is that only random sampling of a class is obtained, since only one student answers each question. This drawback can be overcome by the use of some sort of student response system by the instructor. On a one-to-one basis, as in pre-flight and post-flight briefings, the above is not a problem.

NOTE: Write your answer in the space provided, then compare your answer with the one on the answer sheet on pages 18 and 19.

QUESTION #1

How can oral questions promote mental activity?

QUESTION #2

Why will oral questions maintain student interest during a lesson?

QUESTION #3

What is a drawback in using oral questions to evaluate learning?

Desired Qualities of Good Oral Questions

4. If Oral questions are to serve the purposes stated in paragraph 3, you must be mindful of the following desirable qualities of good questions when composing or preparing to use them.

1. EASILY UNDERSTOOD. Questions should be stated in simple straightforward language; they should be brief, yet complete enough that students have no doubt as to the meaning of the question.
2. COMPOSED OF COMMON WORDS. Questions should be designed to measure knowledge of a subject, not use of language. The use of high sounding words may give you a chance to display your vocabulary, but adds nothing to instruction. Remember, if students do not know the meaning of the

words they will not be able to answer the question. Always keep your vocabulary within the grasp of your student.

3. THOUGHT PROVOKING. Questions should not be so easy that the answer is obvious to all students. Students should be challenged to apply their knowledge. You should avoid using questions where your student has a 50/50 chance of being correct. Examples of these are the YES/NO and TRUE/FALSE type, unless these questions immediately are followed by a "why" or "how" type question.
4. ON MAJOR TEACHING POINTS OF THE LESSON. Questions must be built around the main teaching points of the lessons. They must be asked at the proper place so that these points are emphasized.

Q.3 How to do the scoring of objective and subjective type items? Give examples.

Subjective Assessment

Subjective tests aim to assess areas of students' performance that are complex and qualitative, using questioning which may have more than one correct answer or more ways to express it." Subjective assessments are popular because they typically take less time for teachers to develop, and they offer students the ability to be creative or critical in constructing their answers. Some examples of subjective assessment questions include asking students to:

- Respond with short answers.
- Craft their answers in the form of an essay.
- Define a term, concept, or significant event.
- Respond with a critically thought-out or factually supported opinion.
- Respond to a theoretical scenario.

Subjective assessments are excellent for subjects like writing, reading, art/art history, philosophy, political science, or literature. More specifically, any subject that encourages debate, critical thinking, interpretation of art forms or policies, or applying specific knowledge to real-world scenarios is well-suited for subjective assessment.

Objective Assessment

Objective assessment, on the other hand, is far more exact and subsequently less open to the students' interpretation of concepts or theories. A way of examining in which questions asked has a single correct answer." Mathematics, geography, science, engineering, and computer science are all subjects that rely heavily on objective exams. Some of the most common item types for this style of assessment include:

- Multiple-choice
- True / false
- Matching
- Fill in the Blank
- Assertion and reason

There are two general types of tests:

- **Objective tests** aim to assess a specific part of the learner's knowledge using questions which have a single correct answer.
- **Subjective tests** aim to assess areas of students' performance that are complex and qualitative, using questioning which may have more than one correct answer or more ways to express it

These are some characteristics of objective and subjective tests:

Objective Tests characteristics:

- They are so definite and so clear that a single, definite answer is expected.
- They ensure perfect objectivity in scoring.
- It can be scored objectively and easily.
- It takes less time to answer than an essay test

Subjective Tests Characteristics

- Subjective items are generally easier and less time consuming to construct than are most objective test items
- Different readers can rate identical responses differently, the same reader can rate the same paper differently over time

The "washback or backwash effect is the effect that testing has on the teaching and learning processes.

- The effect can be positive or negative.
- The validity of the testing process can influence the content of our courses, and the way we teach, in a direction that is either with or against the better judgment of the administrators, teachers, students, and parents.

Type of Objective Questions

These are some types of objective question that you can find in tests

1. Multiple- Choice Items
2. True- False Items
3. Matching Items
4. Assertion-Reason Items

Type of Subjective Questions

Subjective questions are questions that require answers in the form of explanations.

Subjective questions include:

1. Essay questions
2. Short answers
3. Definitions
4. Scenario Questions
5. Opinion Questions.

Q.4 Discuss in detail the Goal-Free Evaluation?

In the goal-free evaluation model developed by Michael Scriven (1991), the evaluation looks at a program's actual effect on identified needs. In other words, program goals are not the criteria on which the evaluation is based. Instead, the evaluation examines how and what the program is doing to address needs in the client population. With this model, you observe without a checklist, but record all data accurately and determine their importance and quality. Categories naturally emerge from your observations. This model of evaluation can use all forms of obtrusive methods—those methods a subject is aware of, such as tests—as well as unobtrusive ones—methods that a subject is not aware of, such as a hidden camera—to gather data. The evaluator has no preconceived notions regarding the outcome of the program—that is, goals. The staff should not contaminate the evaluator's method with goal statements. The evaluator is trying to form a description of the program, identify processes accurately, and determine their importance to the program. As the evaluator, you are gathering data on things that are actually happening and evaluating their importance in meeting the needs of the client population.

A good example of this model is the process followed by the Consumer's Union, producers of Consumer Reports, in which the manufacturer's intent for the product is irrelevant to the actual usefulness to the consumer. The goal-free model is the most difficult to use, especially when the evaluator is part of the program or project; yet it is a popular method because it can be used within a program that has many different projects occurring simultaneously. In such a situation the same client population participates in a number of activities, and it is difficult to separate the results of two projects' activities. In fact, program results might come from the interactions between two or more projects' activities.

For example, an evaluator might be asked to evaluate the effectiveness of an adult basic education (ABE) project housed within the program of a local adult learning center (ALC). Also housed in that program are workplace literacy, welfare-to-work, and adult computer literacy projects. Clients of the adult learning center may partake in any or all of these programs. Thus it would be difficult, if not impossible, to isolate the results of just one project's activities. A goal-free evaluation would examine the overall results for the clients of the ALC program, which would be more meaningful than individual evaluations of each project.

The person who performs the goal-free evaluation of the ABE project may have no subject-matter expertise in the field of adult education. This point has become a topic of debate among many experts. Some say the evaluator should have expertise in the field being evaluated; others say no expertise is better (Rossi and Freeman, 1993). The issue, of course, is preconceived notions. Some scholars say that an evaluator who is not familiar with the nuances, ideologies, and standards of a particular professional area will presumably not be biased when observing and collecting data on the activities of a program in that area. They maintain, for example, that a person who is evaluating a program to train dental assistants should not be a person trained in the dental profession. But other scholars allege that a person not aware of the nuances, ideologies, and standards of the dental profession may miss a good deal of what is important to the evaluation. Both sides agree that the

evaluator must attempt to be an unbiased observer and be adept at observation and capable of using multiple data collection methods (Wholey, Hatry, and Newcomer, 1994; 2004).

Once the data have been collected, the evaluator attempts to draw some conclusions about the impact of the program on addressing client needs. This information is then delivered to parties interested in the evaluation results. Again, the evaluator using this model makes a deliberate attempt not to know about program goals, written proposals, or brochures that exist. He or she simply studies the outcomes and reports on them.

The goal-free model works best for qualitative evaluation because the evaluator is looking at actual effects rather than anticipated effects for which quantitative tools have been designed.

Interestingly, Scriven suggests using two goal-free evaluators, each working independently (Popham, 1974). In this way, the evaluation does not rely solely on the observations and interpretations of one person.

As a program manager, you might think that it would be impossible for me to use the goal-free model because I have intimate knowledge of the project and would find it nearly impossible to ignore that knowledge in conducting an evaluation. Similarly, any internal personnel that I might employ to conduct this type of evaluation would also have this knowledge. Consequently, you should probably seek an external, third-party evaluator to perform a goal-free evaluation of your program who has little or no knowledge of the intricacies or nuances of the program to be evaluated.

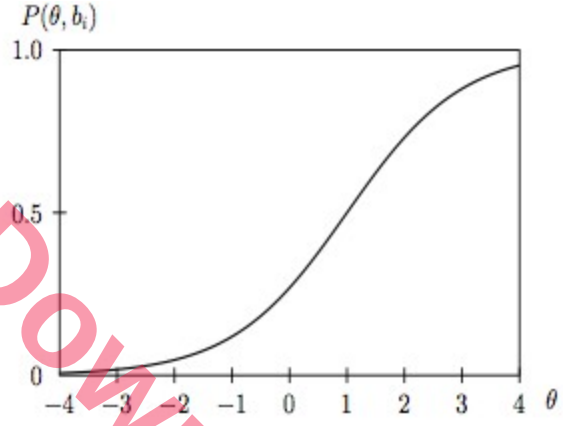
Q.5 Write a comprehensive note on item response theory.

The item response theory (IRT), also known as the latent response theory refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations (i.e. observed outcomes, responses or performance). They establish a link between the properties of items on an instrument, individuals responding to these items and the underlying trait being measured. IRT assumes that the latent construct (e.g. stress, knowledge, attitudes) and items of a measure are organized in an unobservable continuum. Therefore, its main purpose focuses on establishing the individual's position on that continuum.

1) Monotonicity – The assumption indicates that as the trait level is increasing, the probability of a correct response also increases2) Unidimensionality – The model assumes that there is one dominant latent trait being measured and that this trait is the driving force for the responses observed for each item in the measure3) Local Independence – Responses given to the separate items in a test are mutually independent given a certain level of ability.4)Invariance – We are allowed to estimate the item parameters from any position on the item response curve. Accordingly, we can estimate the parameters of an item from any group of subjects who have answered the item.

If the assumptions hold, the differences in observing correct responses between respondents will be due to variation in their latent trait.

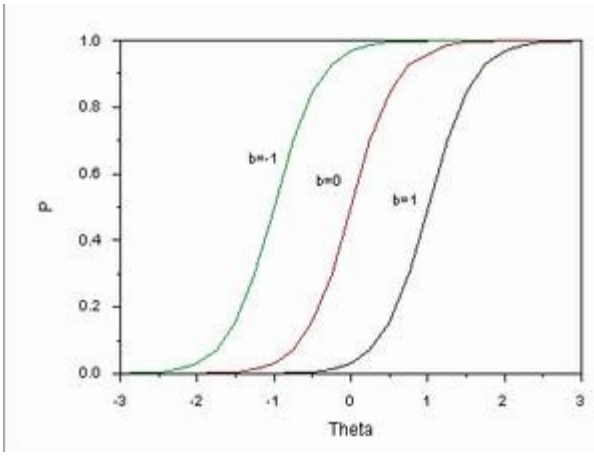
Item Response Function and Item Characteristic Curve (ICC)



IRT models predict respondents' answers to an instrument's items based on their position on the latent trait continuum and the items' characteristics, also known as parameters.Item response function characterizes this association.The underlying assumption is that every response to an item on an instrument provides some inclination about the individual's level of the latent trait or ability. The ability of the person (θ) in simple terms is the probability of endorsing the correct answer for that item.As such, the higher the individual's ability, the higher is the probability of a correct response. This relationship can be depicted graphically and it's known as the Item Characteristic Curve. As is shown in the figure, the curve is S-shaped (Sigmoid/Ogive). Furthermore, the probability of endorsing a correct response monotonically increases as the ability of the respondent becomes higher. It is to be noted that theoretically, ability (θ) ranges from -∞ to +∞, however in applications, it usually ranges between -3 and + 3.

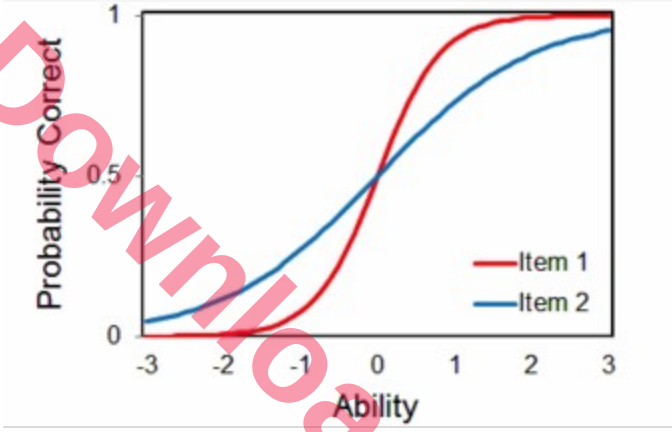
Item Parameters

As people's abilities vary, their position on the latent construct's continuum changes and is determined by the sample of respondents and item parameters. An item must be sensitive enough to rate the respondents within the suggested unobservable continuum.



Item Difficulty (bi) is the parameter that determines the manner of which the item behaves along the ability scale. It is determined at the point of median probability i.e. the ability at which 50% of respondents endorse the correct answer. On an item characteristic curve, items that are difficult to endorse are shifted to the right of the

scale, indicating the higher ability of the respondents who endorse it correctly, while those, which are easier, are more shifted to the left of the ability scale.



Item Discrimination (ai) determines the rate at which the probability of endorsing a correct item changes given ability levels. This parameter is imperative in differentiating between individuals possessing similar levels of the latent construct of interest. The ultimate purpose, for designing a precise measure is to include, items with high discrimination, in order to be able to map individuals along the continuum of the latent trait. On the other hand, researchers should exercise caution if an item is observed to have a negative discrimination because the probability of endorsing the correct answer shouldn't decrease as the respondent's ability increases. Hence, revision of these items should be carried out. The scale for item discrimination, theoretically, ranges from -∞ to +∞ ; and usually doesn't exceed 2; therefore realistically it ranges between (0,2)

Guessing (ci) Item guessing is the third parameter that accounts for guessing on an item. It restricts the probability of endorsing the correct response as the ability approaches -∞.

Population Invariance In simple terms, the item parameters behave similarly in different populations. This is not the case when following the CTT in measurement. As the unit of analysis is the item in IRT, the location of the item (difficulty) can be standardized (undergo linear transformation) across populations and thus items can be easily compared. An important note to add is that even after linear transformation, the parameter estimates derived from two samples will not be identical, the invariance as the name states refers to population invariance and so it applies to item population parameters only.

IRT Model Types

Unidimensional ModelsUnidimensional models predict the ability of items measuring one dominant latent trait.

Dichotomous IRT Models

The dichotomous IRT Models are used when the responses to the items in a measure are dichotomous (i.e. 0,1)

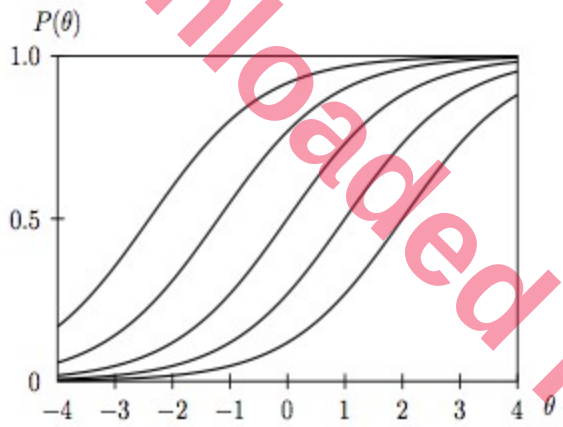
The 1- Parameter logistic model

The model is the simplest form of IRT models. It is comprised of one parameter that describes the latent trait (ability – θ) of the person responding to the items as well as another parameter for the item (difficulty). The following equation represents its mathematical form:

1- PL Model

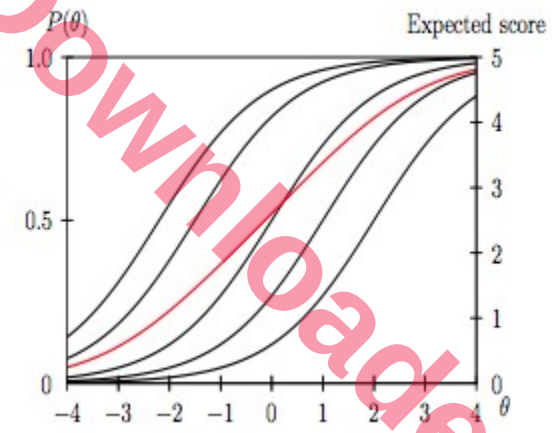
$$P(Y_{is} = 1|\theta_s) = \frac{\exp(1.7a(\theta_s - b_i))}{1 + \exp(1.7a(\theta_s - b_i))}$$

Where,
θ *= ability*
b_i *= difficulty parameter*
a *= Discrimination parameter (fixed and doesn't change between items – no subscript)*
1.7 = scaling factor



The model represents the item response function for the 1 – Parameter Logistic Model predicting the probability of a correct response given the respondent's ability and difficulty of the item. In the 1-PL model, the discrimination parameter is fixed for all items, and accordingly all the Item Characteristic Curves corresponding to the different items in the measure are parallel along the ability scale. The figure shows 5 items, the one on the furthest right is the hardest and would be probably endorsed correctly by those with a higher ability.

Test Information Function



§ It is the sum of probabilities of endorsing the correct answer for all the items in the measure and therefore estimates the expected test score.

§ In this figure, it the red line depicts the joint probability of all 5 items (black)

The Item Information Function

Shows you the amount of information each item provides and it is calculated by multiplying the probability of endorsing a correct response multiplied by the probability of answering incorrectly.

Item Information Function

$$I_i(\theta, b_i) = P_i(\theta, b_i)Q_i(\theta, b_i)$$

Where,
θ = ability
b_i = difficulty parameter
Q = incorrect response
P = correct response

It is to be noted that the amount of information at a given ability level is the inverse of its variance, hence, the larger the amount of information provided by the item, the greater the precision of the measurement. As item information is plotted against ability, a revealing graph depicts the amount of information provided by the item. Items measured with more precision, provide more information and are graphically depicted to be longer and narrower, compared to their counterparts that provide lesser information. The apex of the curve corresponds with the value of bi – the ability at the point of median probability. The maximum amount of information provided would be given when the probability of answering correctly or wrongly are equal, i.e. 50%. Items are most informative among respondents that represent the entire latent continuum and especially among those who have a 50% chance of answering either way.

Estimating Ability

The assumption of local independence, states that item responses should be independent and only associated via

the ability. This allows us to estimate the individual response pattern's likelihood function for the measure administered by multiplication of the item response probabilities. Next, through, an iterative process, the maximum likelihood estimate of ability is calculated. Simply, the maximum likelihood estimate provides us with the expected scores for each individual.

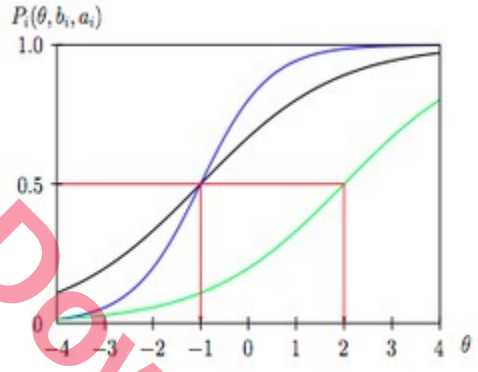
The Rasch Model vs. 1- Parameter Logistic Models

The models are mathematically equal, however, the Rasch Model constrains the Item Discrimination (ai) to 1, while the 1-Parameter logistic model strives to fit the data as much as possible and does not limit the discrimination factor to 1. In the Rasch Model, the model is superior, as it is more concerned with developing the variable that is being used to measure the dimension of interest. Therefore, when constructing an instrument fitting, the Rasch Model would be best, improving the precision of the items.

Rasch Model	$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_j)}$	*Where,* ***θ*** *= ability* b_i *= difficulty parameter*
1- PL Model	$P(Y_{is} = 1\|\theta_s) = \frac{\exp(1.7a(\theta_s - b_i))}{1 + \exp(1.7a(\theta_s - b_i))}$	*Where,* ***θ*** *= ability* b_i *= difficulty parameter* *a = Discrimination parameter (fixed – no subscript)* *1.7 = scaling factor*

The 2- Parameter Logistic Model

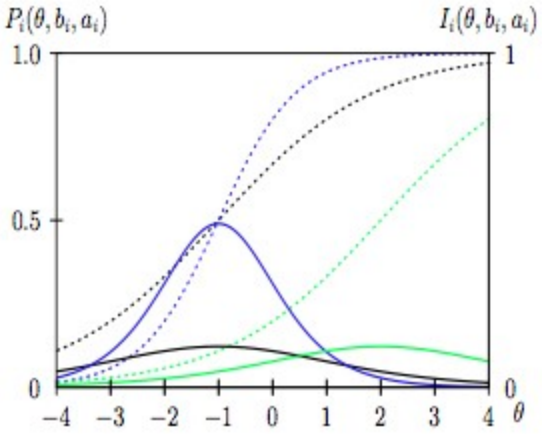
2-PL Model	$P_{ij}(\theta_j, b_i, a_i) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_j)]}.$	*Where,* ***θ*** *= ability* b_i *= difficulty parameter* a_i *= Discrimination parameter (not fixed – can change by item)*



The two parameter logistic model predicts the probability of a successful answer using two parameters (difficulty bi & discrimination ai).

The discrimination parameter is allowed to vary between items. Henceforth, the ICC of the different items can intersect and have different slopes. The steeper the slope, the higher the discrimination of the item, as it will be able to detect subtle differences in the ability of the respondents.

The Item Information Function



As is the case with the 1-PL Model, the information is calculated as the product between the probability of a correct and an incorrect response. However, the product is multiplied by the square of the discrimination parameter. The implication is that, the larger the discrimination parameter, the greater the information provided by the item. As the discriminating factor is allowed to vary between items, the item information function graphs can look different too.

$$I_i(\theta, b_i, a_i) = a_i^2 P_i(\theta, b_i) Q_i(\theta, b_i).$$

Estimating Ability

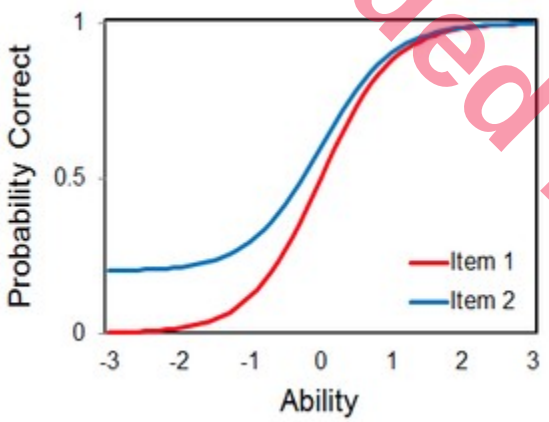
With the 2-PL Model, the assumption of local independence still holds, and the maximum likelihood estimation of the ability, is used. Although, the probabilities for the response patterns are still summed, they are now weighted by the item discrimination factor for each response. Their likelihood functions, therefore, can differ from each other and peak at different levels of θ.

The 3 – Parameter logistic model

3-PL Model

$$P(\theta, a, b, c) = c + (1 - c)\frac{\exp(a(\theta - b))}{1 + \exp(a(\theta - b))}.$$

Where,
θ = ability
b_i = difficulty parameter
a_i = Discrimination parameter (not fixed – can change by item)
c = Guessing parameter (asymptote)



The Model predicts the probability of a correct response, in the same manner as the 1 – PL Model and the 2 PL – Model but it is constrained by a third parameter called the guessing parameter (also known as the pseudo chance parameter), which restricts the probability of endorsing a correct response when the ability of the respondent approaches -∞. As respondents reply to an item by guessing, the amount of information provided by that item decreases and the information item function peaks at a lower level compared to other functions. Additionally, difficulty is no longer demarcated at median probability. Items answered by guessing, indicate that the respondent's ability is lesser than its difficulty.