05. Item Analysis

Item Analysis, a general term for several methods used to evaluate the test item, is one of the most important aspects of test construction. It is the process used in psychology & education, of examining individual test items to evaluate their qualities and effectiveness to determine how well they measure the construct of interest. This basic methods involve assessment of item difficulty and item discriminability.

Purpose

1. **Evaluate the quality of each item :** Item analysis identifies strong items and detects weak or faulty ones that may need revision or removal.
2. **Guide test construction and refinement :** In developing a new test—or modifying an existing one by shortening, lengthening, or improving it—the final set of items is usually selected through systematic item analysis.
3. **Support reliability and validity :** Both the **validity** (measuring what the test is supposed to measure) and **reliability** (consistency of scores) of any test ultimately depend on the quality of its individual items.
4. **Assess item characteristics :** Analysis provides information on:
    - **Difficulty** (how easy or hard an item is)
    - **Discrimination** (how well the item distinguishes between high and low scorers)
    - **Distractor effectiveness** (for multiple-choice items, whether incorrect options work as intended)
5. **Improve measurement and fairness :** Item analysis suggests ways of improving the measurement properties of a test, ensuring items are not biased, ambiguous, or misleading.
6. **Enhance test efficiency :** By removing poor items, tests can be made shorter, more accurate, and less time-consuming without losing quality.

Qualitative vs Quantitative Item Analysis

Aspect	Qualitative Item Analysis	Quantitative Item Analysis
Focus	Content validity (content & form of items)	Item difficulty, item discrimination
Nature	Judgment-based	Statistical
Data	No responses needed	Responsive data is required
When Done	Before test administration	After test administration
Who Does It	Subject experts, teachers, reviewers	Test developers, psychometricians
Purpose	Improves **validity** (measures what it should)	Improves **reliability** (consistency and accuracy of measurement)

Classical Test Theory / True Score Theory

True Score Theory (also known as **Classical Test Theory**, or CTT) is a foundational concept in psychometrics and psychological testing. It helps us understand how test scores relate to the actual ability or trait being measured. According to this theory, if a test were administered multiple times to the same individual under similar construct, the true score would remain constant while the error scores would vary.

Observed Score(X)=True Score(T)+Error(E)

- **Observed Score (X):** The score a person gets on a test.
- **True Score (T):** The person's actual level of ability or trait, if there were no measurement errors.

- **Error (E):** Random factors that cause the score to be higher or lower than the true score (e.g., distractions, fatigue, guessing).

Assumptions

1. **Errors are random** – They are not consistent and normally distributed around true score.
2. Random errors are normally distributed
3. The average of someone's raw score could be the best estimate of their true score.
4. **True score and error are uncorrelated** – Knowing someone's true score says nothing about the size of the error.
5. **Average error is zero** or the mean of distribution score is zero. **(**It means effect is zero not the error. The random errors are just as likely to be positive (making scores a bit too high) as they are to be negative (making scores a bit too low). So, when you average all the errors, they cancel each other out, and the average is zero**.)**

Reliability Coefficient (r)

$$\text{Reliability} = \frac{\text{Variance of True Scores}}{\text{Variance of Observed Scores}}$$

- Ranges from **0 to 1**
- Closer to **1** means **more reliable**

Ramifications (Rules) of CTT

Embretson and Reise (2000) reviewed the ramifications of CTT.

1. **The Standard Error of Measurement of a Test is Consistent Across an Entire Population -** SEM is assumed to be the same for all individuals in a population. High, medium, or low raw scores all share the same SEM. It is derived from a large sample and generalized to the population.
2. **Longer Tests Increase Reliability -** The more items a test has, the more reliable it becomes. Just like surveying more people gives more stable statistics, adding more questions better represents the "universe of items."
3. **Parallel Forms Require Equality -** Multiple forms of a test are considered to be parallel only if they have equal means, variances, reliabilities, and correlations with other variables.
4. **Item Statistics Depend on the Sample -** Difficulty, reliability, etc depend on the representative sample used. Without norms (context of comparison), a test score is meaningless.
5. **True Scores Assumptions -** True scores are assumed to be interval-level data and normally distributed. If not, test developers transform or adjust scores to fit these assumptions.
6. **Test Properties Depend on Original Scale -** If item responses are changed (e.g., a test that had a 4 point Likert-type rating scale for responses now uses a 10 point Likert-type rating scale) then the properties of the test is also altered.

Limitations of CTT

1. **Problems with Difference and Change Scores -** Score changes over time may **not be uniform across all initial levels**. While measuring before and after, change is not equal for all levels. Example: Low scorers improve more than high scorers after training.
2. **Dichotomous Items and Factor Analysis -** CTT suggests dichotomous items (right/wrong) should not be factor analyzed. This limits **validity studies** for many cognitive ability tests as factor analysis is used to check if a test really measures **one main thing** (unidimensional) or **several things** (multidimensional).

3. **Content of Items is Ignored Statistically -** Statistically, CTT only looks at whether participants got it right or wrong, and how the item correlates with total test scores, CTT doesn't statistically analyze the content itself. It ignores wording/content once the test is running.

Item Difficulty

It refers to the proportion of test-takers who answer an item correctly.

$$\text{Item Difficulty} = \frac{\text{No. of students answering the item correctly}}{\text{Total No. of students}}$$

The value of item difficulty (p) ranges from 0 to 1. A higher p-value means easier items and a lower p-value means harder items.

An item with a **p-value of 0 or 1** (i.e., everyone gets it wrong or everyone gets it right) does not help in measuring individual differences, so such items are considered useless. Similarly, too easy ($p > 0.80$) or too difficult ($p < 0.20$) items contribute little to discrimination.

The most useful items are those with **p-values around 0.5**, because they provide the greatest variation in responses and best differentiate between students. Therefore, test constructors aim to include items in this moderate difficulty range and usually remove items with extreme values.

Recommended range for p-value is **.30 and .70**.

Chance Factor

The minimum probability of an item that could be answered correctly by chance alone. For example: True–False question → can be guessed right **50% of the time**. 4-option MCQ → can be guessed right **25% of the time**.

Hence, items must be more difficult than chance performance.

Optimal Item Difficulty

The most effective items are those halfway between perfect performance (1.00) and chance performance.

Example: For a 4-choice MCQ → Chance = .25 ; 100% = 1.00 That means the midpoint = .625 → best difficulty level.

Contextual use

- **Selection tests (**e.g., medical school)**:** More difficult items (to separate top candidates).
- **Special education placement:** More easy items (to separate lower scorers).
- **Classroom tests:** A mix of easy, moderate, and difficult items to measure all students fairly.

 # Human Factor - A few easy items are good for morale and reducing anxiety, even if they don't discriminate well.

Item Discriminability

This assessment determines whether the people who have done well on particular items have also done well on the whole test. How well a **test item** separates people who score **high on the whole test** from those who score **low on the whole test**.

If an item is "good," then: Students who did well overall are **more likely** to answer it correctly, students who did poorly overall are **less likely** to answer it correctly.

Extreme Group Method / Upper Lower Index

This method compares people who have done well with those who have done poorly on a test. The difference between the proportions of people in each group who got each item correct is called the discrimination index.

- **Step 1:** Split test takers into groups:
  - **Top group** (e.g., top 27% of scorers).
  - **Bottom group** (e.g., bottom 27% of scorers).
  - (Sometimes a **middle group** is ignored.)
- **Step 2:** For each item, calculate the **p-value** (proportion correct) for both groups:
  - P upper = proportion of the **top group** who got the item correct.
  - P lower = proportion of the **bottom group** who got the item correct.
- **Step 3:** Compute the **Discrimination Index (D)**:

$$D = p_{\text{upper}} - p_{\text{lower}}$$

- D = 0.4 or more → item is excellent.
- D = 0.3-0.39 → good
- D = 0.11-0.29 → fair
- D = 0-0.1 (around **0)** → poor: item does not discriminate.
- D **negative** → scoring error or flawed

Item	% Correct Top (Pt)	% Correct Bottom (Pb)	di = Pt – Pb
1	.89 (89%)	.34 (34%)	.55 ✅ good
2	.76 (76%)	.36 (36%)	.40 ✅ good
3	.97 (97%)	.45 (45%)	.52 ✅ good
4	.98 (98%)	.95 (95%)	.03 ❌ too easy
5	.56 (56%)	.74 (74%)	–.18 ❌ negative

Item Total Correlation

This assessment checks how much an individual item aligns with the rest of the test. If an item correlates strongly with the total score, it means it is measuring the same construct. If weal or negative, the item may be unreliable.

- **Pearson Product-Moment Correlation** → Used when both variables are continuous (e.g., Likert scale responses and total test scores).
- **Pearson Point- Biserial Correlation**→ Used when one variable is dichotomous (item - right/wrong, 0/1) and the other is continuous (total test score). Most common in test item analysis (multiple-choice items).
- Pearson Biserial Correlation → Used when one variable is false dichotomous (dichotomous with an underlying continuous trait like).

Corrected item total correlation → excluding the item itself from the total score to avoid artificial inflation. While computing correlation between an item and the total including that same item, the result is inflated because the item is part of the total. So as solution the corrected item total correlation is to remove the item from the total score before computing the correlation. This gives a truer estimate of how well the item aligns with the test as a whole.

True Dichotomy → Where the categorization really has only two possible alternatives for a single item (male/female, married/single, yes/no, pass/fail)

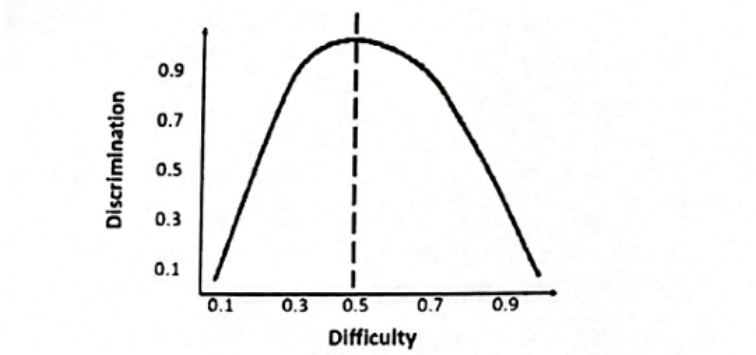
False Dichotomy → Where an arbitrary decision has been made to force a continuous variable into a dichotomous one. For example “you are either smart or stupid - but intelligence isn’t just two extremes.

Relationship between Item difficulty and discrimination

Items of moderate difficulty (p ≈ 0.30–0.70) usually have the highest discrimination. Because they create variability: high scorers get it right, low scorers get it wrong.

Items with **p ≈ 0.5** maximize discrimination. Example: If 100 learners take a test and an item has p = 0.5 → about 50 students get it right, 50 get it wrong. This provides the **highest spread** for distinguishing ability levels.

- **Very easy items (p > 0.90)** → everyone gets them right, so they don't distinguish between high and low scorers.
- **Very hard items (p < 0.10)** → almost no one gets them right, so again, they don't discriminate well.



This graph shows an inverted U-shaped curve. Discrimination is highest at medium difficulty. Discrimination decreases when items are too easy or too hard.

Item Response Theory

A modern measurement paradigm used for analyzing test items, designed for categorical data. Unlike CTT, which focuses on total test scores, IRT emphasizes the relationship between the latent trait (unobservable ability/skill, θ) and the observed responses (correct/incorrect answers, item scores)

IRT models the **probability** that a person with a certain ability level will answer an item correctly, based on characteristics of the **item**. The chance (probability) that someone answers a question right depends on how skilled they are and how the question behaves (like how hard it is, how tricky it is, or if guessing might help). A test item isn't just "hard" or "easy" in general — it behaves differently depending on who's answering it. Your score shouldn't just be the total number of right answers — it should reflect which items you got right and how much they tell us about your ability.

The Three Item Parameters:

1. **Item Difficulty / Location Parameter (b)**
    - Represents the ability (θ) at which a person has a 50% probability of endorsing the item. The higher b the more difficult item requiring higher ability.
    - Analogous to difficulty index in Classical Test Theory.
    - Ranges typically from -3 to +3 like z-scores.
2. **Item Discrimination Parameter (a)** –
    - Refers to the slope of the Item Characteristic Curve at point b.
    - Indicates how sharply the item can differentiate high vs. low ability people → Higher a = better at distinguishing.
    - Similar to factor loadings in factor analysis
3. **Guessing Parameter(c)** –

- Accounts for the possibility that low-ability examinees may guess it right and answer correct by chance.
- Most relevant in multiple-choice tests.
- Typical values : Around the reciprocal of number of options (e.g., 0.25 for 4-option mcq)

Assumptions

1. **Unidimensionality -** The test measures **a single underlying latent trait (θ)** (e.g., math ability, reading skill). All items are assumed to reflect this **one dominant ability**.
2. **Local Independence -** Once the latent trait (θ) is controlled for, item responses are independent. In other words: A student's response to one item should not directly affect their response to another. Any correlation between items should be explained **only** by the latent trait.
3. **Monotonicity -** The probability of answering an item correctly should **increase monotonically** as the latent trait (θ) increases. Example: A more able learner should have a higher chance of success on every item than a less able learner.
4. **Item Characteristic Curve (ICC) Validity -** Each item has a specific mathematical function (logistic curve) relating θ to the probability of a correct response. Assumes that this functional form accurately represents how items behave.

Advantages / Importance

1. **Focuses on item-level analysis, not just the test score**
   - You could get fewer items right than someone else but still have a higher ability level if your items were harder.
2. **Tests can be adaptive** (Computer Adaptive Testing, CAT)
   - The computer gives questions suited to your ability.
   - Saves time: no need to answer too-easy or too-hard questions.
   - Reduces cheating: different students get different items.
3. **Precision at all ability levels**
   - Conventional tests often measure best around average ability.
   - Adaptive IRT tests keep high precision for low, average, and high ability students.
4. **Handles different item formats**
   - Multiple choice, Likert scale, true/false, etc.
   - Can detect unusual response patterns or inattentiveness.
5. **Allows identification**
   - Identifies which items are too easy, too hard, or discriminatory.

Aspect	**Item Analysis in CTT**	**Item Analysis in IRT**
Focus	Overall test scores and item performance	Item-level performance across ability levels
Item Difficulty	Proportion of students who answered correctly (p-value)	Location parameter (b) → ability level where 50% chance of correct
Item Discrimination	Correlation between item score & total test score	Discrimination parameter (a) → slope of item curve
Guessing	Not directly considered (sometimes implied)	Guessing parameter (c) → lower bound of chance success

Aspect	**Item Analysis in CTT**	**Item Analysis in IRT**
Population Dependence	Item statistics depend on the sample of test-takers	Item parameters are (theoretically) sample-independent
Application	Simple test construction, classroom exams	Advanced testing (e.g., GRE, CAT, large-scale assessments)
Output	Item difficulty index, discrimination index	Item Characteristic Curve (ICC), detailed parameters (a, b, c)