1. **Norm-Referenced Testing**

The goal is to differentiate among students, ranking them from high to low in a bell-shaped curve. Primary use in placement test or diagnostic tests (NOT in classroom grading).

FIRST CHECK WETHER DATA DICHOTOMOUS OR POLYTOMOUS

* + - 1. Item Difficulty (item facility) – estimate of item difficulty

1. IF = The proportion of test takers who answered a given question correctly (0.0 – none, 1.0 – all, for dichotomous – 0 or 1)
2. IF\* = for polytomous (IF divided by the number of points possible for that item). The scale is the same (0 -1).

The ideal IF is around 0.50. If achieved, it will provide the max info about for the largest group of takers (middle of the normal distribution). However, usually used 0.30 and 0.70. IF below 0.30 is too difficult, above 0.70 - too easy.

* + - 1. Item Discrimination – how well a given item separates examinees with high and low ability.

(how student did on a item VS their overall scores on the test containing that item).

If large data set with normal distribution – correlational.

1) Correlational approach

- Point-biserial (r(pb)) - a dichotomous vs interval (Ex. Measure the relationship between how students performed on one test question and on the overall test).

- Pearson r - a polytomous vs interval

If multiple sections (listening and writing, for example), discrimination is based on section total.

2) Subtractive approach - ID or ID\*

IF of top scored group – IF of low scored group. Allow for 25%, 27%, 33% size groups depending on the size of the group tested. IF upper- IF lower

* + - 1. Rules of Thumb for interpreting NRT Item Analysis

To help to ensure satisfactory reliability. Reliability is affected by the number of items on a test and by the quality of each item.

For high-stakes test the best results are:

|  |  |
| --- | --- |
| Approach | Guideline |
| Difficulty  Discrimination (the higher the better)  Correlational approach (r(pb)  Subtractive approach (IDul) | 0.30 <= IF <= 0.70 (target 0.50)  R(pb) >= 0.30  IDul >= 0.40 |

For teacher-made tests:

|  |  |
| --- | --- |
| Approach | Guideline |
| Difficulty  Discrimination (the higher the better)  Correlational approach (r(pb)  Subtractive approach (IDul) | 0.30 <= IF <= 0.70 (target 0.50)  R(pb) >= 0.15  IDul >= 0.20 |

Low reliability: the test is too short or/and the items are poorly discriminated.

Near 0 discrimination: IF = 1.0 (too difficult) or IF = 0.0 (too easy). Meaning i8t contributes nothing to the reliability of the test.

Negative discrimination: harming the reliability level. -> Check the scoring key. If the key is correct, the item needs reviewing (too ambiguous, too deceptive, two correct keys etc.)

1. **Criterion-Referenced Testing**
2. Item Difficulty

Similarly 0.50 is the best. But as it is difficult, IF is only used to interpret other results and can’t be used for keeping or rejecting an item. For example, pre-test CRT – negatively skewed, post-test CRT – positively skewed.

1. Item Discrimination – item performance VS mastery/non-mastery

1) Correlational approach

- Tetrachorus cor coeff - a dichotomous vs interval (mastery/non-mastery, pass/fail etc.)

- Tetrachorus cor coeff and Point-biserial – if scores are not yet set or trusted

- Pearson r - a polytomous vs interval

2) Subtractive approach

- Difference Index (DI) - IF masters – IF non-masters

* Differential groups studies: two existing group are identified: masters and non-masters.
* Intervention studies: two groups of the same students: before instructions and after. Pre-test and post-test model.

- B-index (B) - IF masters – IF non-masters

When multiple cut scores, more groups needed (not only 2). So, the index is calculated for each score separately relatively to the other groups (since changing the cut score changes who is a master or non-master.

1. Rules of Thumb for interpreting NRT Item Analysis

|  |  |
| --- | --- |
| Approach | Guideline |
| Difficulty  Discrimination (the higher the better)  Correlational approach (Tetrachorus cor coeff)  Subtractive approach (DI or B-index) | N/A (if reported only to help discrimination estimates)  Phi >= 0.30  DI >= 0.40  B >= 0.40 |