FAQ Exam Analysis

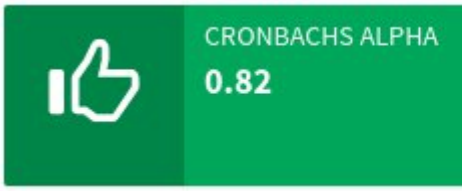
19.7.2024

Classical Test Theory (CTT)

Classical test theory analyzes test results using "classical statistics". The assumption here is that each observed score of an individual consists of a true value and an error value. The true score represents the actual measure of the ability or characteristic that the test is intended to measure, while the error score reflects random influences and measurement inaccuracies. The KTT aims to assess the reliability and validity of tests by examining the consistency of measurement results (reliability) and the accuracy with which a test measures what it intends to measure (validity).

Cronbachs Alpha

Cronbachs Alpha is a measure of the internal consistency or reliability of a test or questionnaire that indicates how closely the individual items are related to each other. It is often used to assess the homogeneity of items and to ensure that all questions measure the same thing. A high value of Cronbach's alpha indicates that the items are strongly related to each other and therefore have a high internal consistency. Values of Cronbach's alpha range from 0 to 1, with higher values indicating better reliability.

		
0.7 < : bad	0.7 < x < 0.8 : acceptable	> 0.8 : good

Item difficulty: The ratio of the number of test participants who answered the item correctly to the total number of test participants. A simple example: If 80 out of 100 participants answer an item correctly, the item difficulty is 0.8 (or 80%).

The item difficulty (p_i) is calculated as followed:

$$p_i = \frac{R}{N}$$

with:

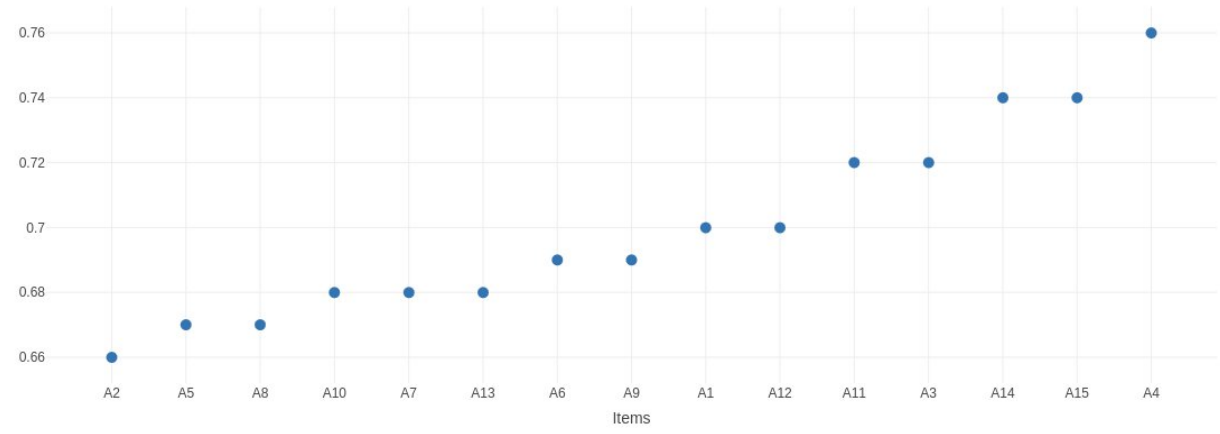
- p_i the item difficulty of Item i,
- R the number of test participants who answered the item correctly,
- N the total number of test participants.

A value of $p=1$ means that all participants answered the item correctly (easy item), while a value of $p=0$ means that no one answered the item correctly (difficult item). Ideally, items should have a medium difficulty (close to 0.5) to allow good differentiation between test participants.

Shown below on a scatterplot:

x-axis : item name

y-axis : item difficulty



Item-Respose Theorie (IRT)

Item response theory is a psychometric model used to analyze test data and questionnaires. It provides a method for evaluating the relationship between the characteristics of the test participants and the characteristics of the test items. In contrast to classical test theory, IRT considers the probability that a person will answer a particular item correctly as a function of their ability and the characteristics of the item.

Item difficulty: Item difficulty is defined as the point on the ability scale at which the probability of answering the item correctly is 50%. This is described by the threshold parameter (b).

A frequently used IRT model is the Rasch model. This model is also used here. The formula for calculating the probability (p) of a correct answer to an item i by a person j with the ability Θ_j is as follows:

$$P(X_{ij} = 1 \mid \Theta_j, b_i) = \frac{1}{1 + e^{-(\Theta j - bi)}}$$

$P(X_{ij} = 1 \mid \Theta_j, b_i)$: the probability that person j answers item i correctly

Θ_j : the ability of the person j ,

b_i : the difficulty of the item i .

In this model, item difficulty (b_i) is the point at which $\Theta_j = b_i$, meaning that the person's ability is equal to the difficulty of the item. At this point, the probability of a correct answer is 50%.

Wright-Map

A Wright map shows the abilities of the test participants and the difficulties of the test items on the same scale. This enables a clear visualization and direct comparison of the positions of individuals and items in relation to the underlying ability or the characteristic to be measured.

- On the Wright map, the abilities of the test participants are shown on the left-hand side of the axis (person axis) and the difficulties of the test items on the right-hand side of the axis (item axis).
- Each point on the person axis represents the ability of a test taker, while each point on the item axis represents the difficulty of an item.
- The scale describes the probability with which a person answers an item correctly.

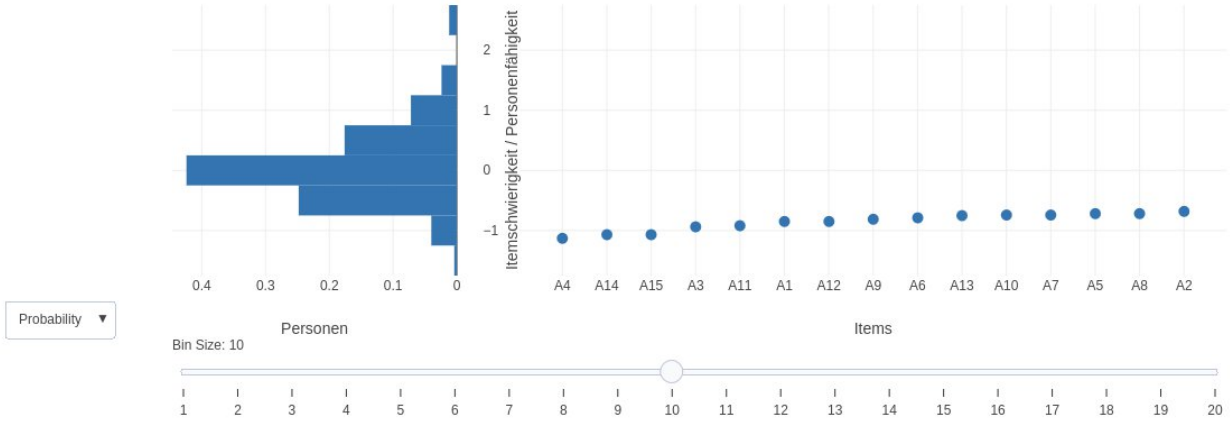
A well-designed Wright map is a valuable tool for assessing the fit between test items and test takers and helps to improve the quality and suitability of a test.

Item axis : Item difficulty (IRT)

Person axis : Person ability

Bin-Size : Setting the bins of the person capability

Persons can be displayed by frequency or number



Good?	*Bad?*
Clear and detailed presentation of the distribution of test takers' abilities and the difficulty of the test items	Uneven distribution of items along the ability scale, e.g. when most items are concentrated in a particular difficulty range, resulting in poor differentiation.
Balanced distribution of items along the ability scale. -> Test contains items for different levels of difficulty and is therefore suitable for test takers with different levels of ability	Test taker ability and item difficulty are not well matched e.g. if many items are too difficult or too easy for the majority of test takers, the map is also poor.
Items differentiate well and are neither too easy nor too difficult, which enables optimal measurement of skills	

Histogram

Graphically displays the distribution of item difficulties in a test or questionnaire. It shows how often different levels of difficulty of test items occur and makes it possible to visually record the general distribution of difficulty ratings.

- **Axes:** The difficulty levels of the items are plotted on the horizontal axis (x-axis), while the vertical axis (y-axis) shows the frequency or number of items that fall within a certain difficulty range.

- **Bars:** Each bar represents the number of items that lie within a certain difficulty range. The height of a bar indicates how many items are in this range

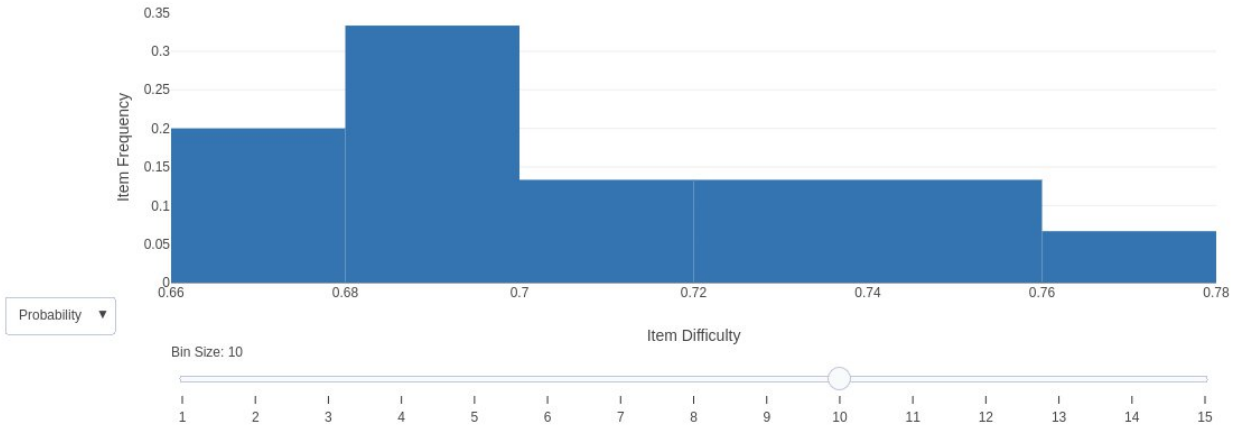
A well-designed histogram enables effective analysis and helps to evaluate the quality of the test by providing insight into the distribution of item difficulties and identifying possible adjustments for better test coverage.

x-axis : item difficulty

y-axis : Item frequency

Bin-Size : Setting the bins of the item difficulty

Item difficulty can be displayed according to frequency or number



Good?	Bad?
Good Distribution: The distribution of item difficulties is balanced. Ideally, the difficulty levels of the items should cover a wide range, including both easy and difficult items. This ensures that the test is suitable for all test takers with different levels of ability.	**Inconsistend Distribution**: Distribution of item difficulties is highly unbalanced. For example, if all items are concentrated in a narrow difficulty range, which indicates that the test contains either very easy or very difficult items.
No Gaps: No large gaps seen in the difficulty ranges, indicating that the test has comprehensive coverage of the entire difficulty range.	**Coarse granularity:** If the difficulty levels are grouped in too coarse intervals, this can obscure important details about the actual distribution of item difficulties.

Violin-Plot

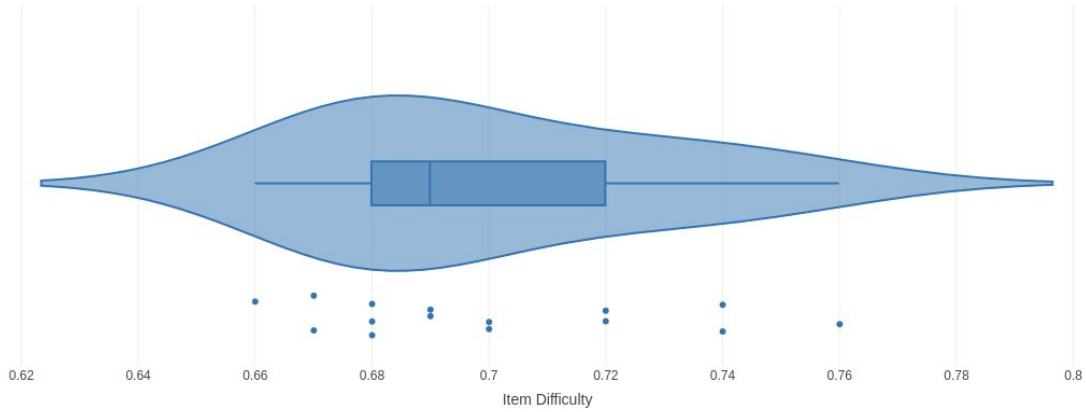
Visualizes the distribution of difficulty ratings for different test items. It combines the features of a box plot and a density plot by displaying the distribution of the data on both sides of a vertical axis as a violin.

- The width of the "violin" at a particular location indicates the density of data points at that location. Wider sections mean that more items have a similar difficulty, while narrower sections indicate that fewer items share this difficulty

- In addition to the usual boxplot statistics (such as median, quartiles and outliers), the violin plots provide a detailed representation of the distribution of item difficulties, including possible multi-peakedness or asymmetric distributions.

A good violin plot helps to identify patterns in item difficulty, such as clusters of similarly difficult items or differences in difficulty across different item groups. A poor violin plot, on the other hand, can be misleading and miss important details of the distribution.

x-axis : item difficulty



ViolinPlot with Boxplot

Good?	Bad?
A clear visualization of the density of data points	Too few data points, resulting in inaccurate density estimation
An easily recognizable central tendency and variability of item difficulties	Excessive smoothing of the density function, which obscures important details of the distribution
No excessive bias or misinterpretation of the data	Missing or misleading labels and axis scaling that make interpretation difficult.

Factor analysis

Factor analysis is used to identify the underlying dimensions (factors) of a data set by examining the relationships between the variables. In the context of item difficulty, factor analysis aims to find out whether there are common dimensions or groups that explain the difficulty of different items in a test. This can help to understand the structure of a test and assess whether the items together measure a consistent construct.

A good factor analysis shows a clear and interpretable structure, high factor loadings and high commonality. Poor factor analysis results in unclear or difficult to interpret factors and low loadings.

What are Factors?

Factors are latent variables or dimensions that explain the observed variables (e.g. item difficulties). They represent underlying constructs or patterns in the data.

What is a MR (Minimal Residual)?

Minimal residual refers to the method for estimating the factor loading in the factor analysis. The method minimizes the residuals (differences between the observed correlation matrices and the correlation matrices estimated by the model) to determine the best fit.

What is R^2?

is a measure of the proportion of the variance of a variable that is explained by the factors. It indicates how well the factors explain the variability of the item difficulties. A high R^2 means that a large proportion of the variance is explained by the identified factors, which indicates a good model fit.

What is a rotation?

Rotation is a procedure for simplifying and improving the interpretability of the factors. There are two main types of rotation: orthogonal rotation (e.g. Varimax) and oblique rotation (e.g. Promax). Orthogonal rotation leads to uncorrelated factors, while oblique rotation allows for correlated factors.

Gut?	Schlecht?
Clarity of structure of factors: Clear and interpretable structure of factors. This means that the items are well grouped based on their levels of difficulty and that these groups (factors) are logical and theoretically meaningful.	**Unclear structure of factores:** Unclear or difficult to interpret factors. This can happen if items have high loadings on several factors or if there is no clear grouping of the items.

High factor loadings: Items should have high loadings on only a few factors, which means that they correlate strongly with the factors on which they are placed. This shows a good distinction between the factors and the items.	**Low factor loadings:** If items have low loadings on the identified factors, this indicates that the items do not fit the factors well and may not measure the underlying construct well.
Good communality: Items sollten eine hohe Kommunalität aufweisen, d.h. der Großteil der Varianz eines Items sollte durch die identifizierten Faktoren erklärt werden.	**High cross charges:** When items have high loadings on multiple factors, this can lead to confusion and make it difficult to interpret the factors.