# Item scaling

Classical item analysis and Rasch measurement procedures are often combined in practice to identify good items for a test. We should combine the following CTT statistics when using IRT.

(a) identify items with negative item-total correlations (i.e. negative classical item discrimination),

(b) evaluate outfit before infit,

(c) examine mean square fit statistics before standardized fit statistics,

(d) focus on high values before low values.

Item selection procedures in classical test theory tend to focus on item discrimination, whereas item selection procedures in Rasch measurement emphasize item and person fit.

I have analyzed items here according to CTT. Later i have used MIRT to analyze the items. I used the **Environment** dataset in ltm package . I used the following code to recode the responses in numbers and then i extracted the dataset as an excel file. In the analysis i have not yet deleted or excluded any items yet as this is just a demo. I have also added a lot of theory to aid in interpretation of various statistics.

There is no missing data. If there were , i could have used pairwise or listwise deletion. I could also have used various imputation methods such as a series mean, linear interpolation , regression imputation or a full information maximum likelihood approach. In IRT , missing data is not usually an issue as IRT can handle missing data pretty well.

I used the following code to recode my qualitative responses into quantitative numbers.

```
data(package = "ltm")

library(tidyverse)

df<- Environment

x<-df %>% mutate_all(~ str_replace(., "^$", NA_character_)) %>% mutate_all(.funs = ~ as.integer(recode(.x = .,"not very concerned"=0,"slightly concerned"=1, "very concerned"=2)))

library("writexl"

write_xlsx(df,"D:\\df.xlsx")
```

df %>% mutate_all(~ str_replace(.,"^$", NA_character_)) %>%mutate_all(.funs = ~ as.integer(recode(.x = .,"not very concerned"=0,"slightly concerned"=1, "very concerned"=2)))

## Scale score transformation

Raw scores are easy to compute, but the raw score scale is vulnerable to test modifications that affect test difficulty and the distribution of scores. The raw score scale may change whenever you add, remove, or replace items on a test. For example, a test composed of 10 binary items has a raw score scale that ranges from 0 to 10, but adding 2 items to the test changes the scale to range from 0 to 12. There is no way to compare scores from the 12-item test to those from the 10-item test without making any adjustments to the scale. Even if the number of items remains the same, replacing difficult items with easier ones will affect the score distribution. Examinees will appear to earn higher scores while in reality the test is easier. Multiple versions of the same test also involve differences in test difficulty and scores that are not directly comparable. Although differences in test difficulty and the distribution of scores may be overcome with test equating, the consistency of score interpretation is achieved through the use of a common score scale and the conversion of raw scores to scale scores.

We can use percentile scores, normal scores and normalized scores to scale the data

There are many types of normal scores such as van der Waerden's normal score, Bloom's normal score ,Tukey's normal score. Normal scores are commonly encountered in nonparametric statistics such as nonparametric item response theory. We rarely use them as test scores in an operational testing program.

Normalized scores are percentile scores from the normal distribution. Converting raw scores to normalized scores or normal scores forces the data to be normally distributed (i.e. bell shaped). If the raw score distribution is symmetric and unimodal, then this conversion is a good approximation. On the other hand, if the raw score distribution is substantially skewed, then normalized scores or normal scores are only a rough approximation of the raw score distribution.

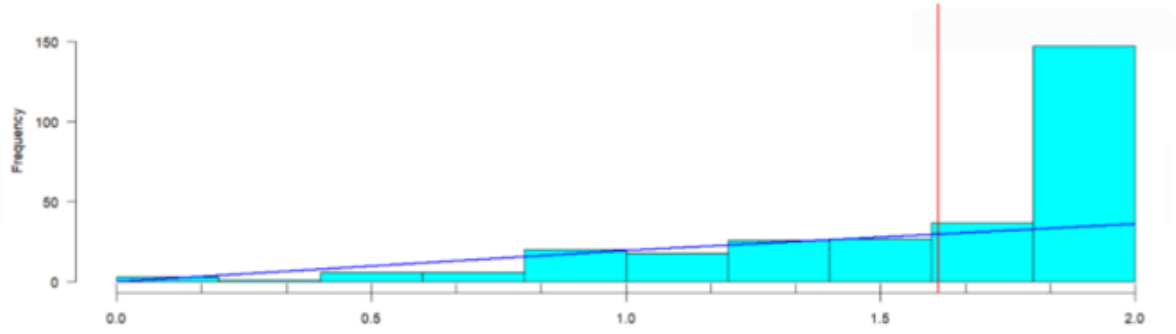
I transformed the scores with mean of 3 and SD of 1. I also rounded off to nearest Integer. The scores are also scaled as percentile and normalized scores. These scores can be helpful in deciding cut off scores. More appropriate ways of deciding cut off scores include the Angoff method and its variations.

I plotted the data and it seemed heavily skewed. Normal scores should not be used with this data. Percentile ranks are also not suitable scores for all testing applications. You need a sample size of several thousand examinees to obtain stable percentile ranks. We do not have enough sample size to use percentile scores accurately either.

SCORE TABLE

| Original Value | Percentile Rank | Normalized Score |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| 3 | 2 | 1 |
| 4 | 4 | 1 |
| 5 | 7 | 2 |
| 6 | 10 | 2 |
| 7 | 15 | 2 |
| 8 | 23 | 2 |
| 9 | 32 | 3 |
| 10 | 43 | 3 |
| 11 | 58 | 3 |
| 12 | 84 | 4 |

Linear Transformation: Y = 1.00X + 3.00



Sum Score Descriptive Statistics

| Statistic | Value |
|---|---|
| N | 291.0000 |
| Min | 0.0000 |
| Max | 12.0000 |
| Mean | 9.6838 |
| St. Dev. | 2.5844 |
| Skewness | -1.2789 |
| Kurtosis | 1.2953 |

Normalized Score Descriptives

| Statistic | Value |
|---|---|
| N | 291.0000 |
| Min | 0.0000 |
| Max | 4.0000 |
| Mean | 2.9931 |
| St. Dev. | 0.8982 |
| Skewness | -0.6189 |
| Kurtosis | -0.0983 |

These statistics could have been used to create various norms if normative interpretation was needed. A norm is the distribution of scores for a representative sample of examinees from the target population. It is characterized by statistics such as the mean, standard deviation, and percentile ranks . Norm-referenced scores such as percentile ranks and normalized scores allow for relative comparisons among examinees and permit statements such as “Jill scored above average” or “Trevor performed better than 86% of examinees.” It is through the norm (i.e. the performance of others) that a score takes on meaning. However, the interpretation of a norm-referenced score is not absolute. It is relative to the target population. Norms can be created for different populations. A national norm represents the distribution of scores from a nationally representative sample of examines, and a local norm may be the distribution of student scores within a school district. Because the interpretation is relative, a low score on a national norm may be a high score on a local norm. If you change the norm, you also change the interpretation of scores.

Criterion-referenced scores may be expressed as the percentage of items answered correctly or as a statement about the student’s achievement level (e.g. fail, pass, or high pass). Percentage correct scores could have been used for criterion reference testing.

Although many tests can be classified as either norm or criterion referenced, it is not uncommon to find a test with both frames of reference. Educational tests often report a pass/fail score along with a percentile rank. The goal is to not only identify students who have mastered the tested content but also make finer distinctions among examinees in terms of their performance.. The pass/fail categorization of an examinee's performance is a criterion-referenced score, but a percentile rank of, say, 82 is a norm-referenced score. There is no analysis you can conduct that will produce the frames of reference. You must collect data from a representative sample from the population in a norm-referenced test, and you must have a well-defined content domain for a criterion-referenced test.

## Item analysis

The difficulty states the item difficulty according to classical theory analysis. The closer the number is to 2 the easier it is. Dividing each of these values by the maximum possible item score puts item difficulty back onto a 0 to 1 scale. The fourth option(SD) is useful when selectingitem with the largest amount of variance.

ITEM ANALYSIS
qc.DFF
April 19, 2021 23:25:30

| Item | Option (Score) | Difficulty | Std. Dev. | Discrimin. |
|---|---|---|---|---|
| leadpetrol | Overall | 1.5567 | 0.6041 | 0.4517 |
| | 0.0(0.0) | 0.0584 | 0.2349 | -0.4469 |
| | 1.0(1.0) | 0.3265 | 0.4697 | -0.5647 |
| | 2.0(2.0) | 0.6151 | 0.4874 | 0.4554 |
| riversea | Overall | 1.7766 | 0.4715 | 0.5727 |
| | 0.0(0.0) | 0.0241 | 0.1535 | -0.4874 |
| | 1.0(1.0) | 0.1753 | 0.3808 | -0.5962 |
| | 2.0(2.0) | 0.8007 | 0.4002 | 0.5368 |
| radiowaste | Overall | 1.6838 | 0.5840 | 0.6645 |
| | 0.0(0.0) | 0.0619 | 0.2413 | -0.6243 |
| | 1.0(1.0) | 0.1924 | 0.3949 | -0.5713 |
| | 2.0(2.0) | 0.7457 | 0.4362 | 0.6383 |
| airpollution | Overall | 1.6186 | 0.5467 | 0.6803 |
| | 0.0(0.0) | 0.0309 | 0.1734 | -0.5275 |
| | 1.0(1.0) | 0.3196 | 0.4671 | -0.6701 |
| | 2.0(2.0) | 0.6495 | 0.4780 | 0.6172 |
| chemicals | Overall | 1.6907 | 0.5759 | 0.6550 |
| | 0.0(0.0) | 0.0584 | 0.2349 | -0.5769 |
| | 1.0(1.0) | 0.1924 | 0.3949 | -0.6142 |
| | 2.0(2.0) | 0.7491 | 0.4343 | 0.6560 |
| nuclear | Overall | 1.3574 | 0.7401 | 0.5630 |
| | 0.0(0.0) | 0.1581 | 0.3654 | -0.6953 |
| | 1.0(1.0) | 0.3265 | 0.4697 | -0.3695 |
| | 2.0(2.0) | 0.5155 | 0.5006 | 0.5099 |

Item difficulty and discrimination are commonly used in conjunction when selecting items for a norm-referenced test. Item discrimination values should be between 0.3 and 0.7, and binary items should have a difficulty value near 0.5. If guessing is a factor, then the ideal difficulty value is slightly more than halfway between chance and 1 . For example, a multiple-choice item with four options has an ideal difficulty value that is slightly more than [0.25 − (1 − 0.25)/2] = 0.625. These rules of thumb are based on the idea that these values maximize item variance and subsequently increase score reliability. Of course, every item will not likely have an item difficulty that is exactly the ideal value, but items with a difficulty within 0.2(within 0.5) points would be acceptable. For example, for a multiple-choice question in which guessing is not a factor, difficulty values between 0.3 and 0.7 would be acceptably close to the ideal value of 0.5. This range of values will maximize information that the test provides about score differences . No comparable rules of thumb exist for polytomous items, but by extension, it is best to select polytomous items that have moderate discrimination values and large item variances.

Item difficulty is the mean item score. For multiple-choice, true/false, and other items that are scored as right (1 point) or wrong (0 points), **item difficulty is the proportion of examinees who answered the item correctly**. It ranges from 0 to 1, and, despite the name “item difficulty,” a large value of this statistic indicates an easy item and a small value indicates a hard item. For example, an item difficulty of 0.8 indicates that 80% of examinees answer the item correctly. On the other hand, an item difficulty of 0.1 shows that only 10% of examinees answer the item correctly. An item with a difficulty of 0.1 is much more difficult than one with a difficulty of 0.8.

Item difficulty for a polytomous item, an item scored in more than two ordinal categories, is simply the item mean or average item score. It ranges between the minimum possible item score and the maximum possible item score. Interpretation of the item difficulty for a polytomous item depends on the minimum and maximum possible item scores. Therefore, a 4-point polytomous item scored as 0, 1, 2, and 3 points has a mean that ranges between 0 and 3 points. **The closer the mean is to 0, the more difficult the item (e.g. the harder to achieve the highest category), and the closer the mean is to 3, the easier the item (e.g. the easier it is to attain the highest category).** It is possible to **convert the item difficulty for a polytomous item into a proportion correct score by dividing the item mean by the maximum possible item score.**

For example, suppose an item is scored as 0, 1, 2, and 3 points. An item mean of 2.35 corresponds to a proportion correct score of 2.35/3 = 0.78. As with binary items, when the polytomous item difficulty is **converted to a 0 – 1 metric, values close to 1 indicate an easy item, and values close to 0 a difficult item.** However, it would **not be safe to say that 78% of examinees answered this item correctly. This conversion only allows us see that, on average, examinees obtained a large portion of the available points**

The term “item difficulty” makes sense in educational measurement where questions have a correct answer or can be rated on degrees of correctness. In psychological research where the goal is to measure an attitude or personality characteristic, the term is less appropriate. Psychological measures often involve Likert scales that ask people to respond as “Strongly Agree,” “Agee,” “Disagree,” or “Strongly Disagree.” There is no correct answer to such a

question. Any response is acceptable, and the term “item difficulty” does not apply. For Likert and similar types of items, you can think of the item mean as an index of **item endorsability—the extent to which the highest response option is endorsed.** This term is more consistent with the notion of someone endorsing a particular attitude, where all attitudes are acceptable and none are “correct.”

Item discrimination is the extent to which an item differentiates between examinees that obtain different scores on the test. If discrimination is **high**, the item can easily distinguish between **examinees that have similar, but not identical,** test scores. On the other hand, if discrimination is low, the item can **only distinguish** between examinees that have **very different test scores**. An easy way to quantify discrimination is the D-index. For binary items, it is computed as the difference in item difficulty for the top 27% and bottom 27% of examinees. Large values indicate an item is much easier for top-scoring examinees than it is for low-scoring examinees. The main limitation of the D-index is that it does not make use of all available data. The middle 46% of examinees are eliminated from the computation. I used item discrimination statistics that retain all of the available data.

**Item discrimination** is a correlation between the item score and the total test score. It is often called the **item-total correlation** for this reason. **Pearson’s correlation** is the most basic type of correlation involved in an item analysis, and it can be applied to binary and polytomous items. It has a different name when it is computed between a binary item and the total test score. In this situation, it is more specifically referred to as a point-biserial correlation. All item-total correlations have the same interpretation. **Positive item-total correlations mean that high-scoring examinees tend to get the item correct, and low-scoring examinees tend to get it wrong. High positive values for the item total correlation indicate a large amount of discrimination.** Values near 0 reflect little to no discrimination. An item-total correlation can take on negative values, but such a result would indicate a problem. It would mean that low-scoring examinees tend to get the item correct. If you ever observe negative item discrimination, check the item itself and the item scoring. You may have provided the wrong answer code, or it could be a reverse-worded item. **If the answer key is correct, a negative item discrimination value indicates a serious problem with the item.**

Distractor Analysis The proportion of examinees endorsing a distractor provides information about the plausibility of the distractor. Large values indicate that the distractor attracts many examinees, and small values indicate that few examinees selected the distractor. Proportions close to 0 suggest that the distractor is not functioning, and it is possible to eliminate it as a choice. Distractor proportions close to 1 may indicate an item that is excessively difficult, or it may be a sign of an incorrect answer key. Distractor-total correlations show the relationship between a distractor and the total test score. They are computed by coding the distractor as 1 point for selected and 0 points for not selected and then correlating this binary distractor score with the total test score. It is either a Pearson or polyserial type correlation, depending on the option you selected for the correlation. We expect to find negative distractor-total correlations because examinees earning higher scores should be less likely to select a distractor. Stated differently, examinees that select a distractor should earn lower overall test scores. The ideal pattern for a multiple-choice question is to have a positive item discrimination and negative

distractor-total correlations. Any deviation from this pattern should be examined closely. Possible explanations for a positive distractor total correlation include a mistake with the answer key and a distractor that is a legitimate answer. The former can be verified by a review of the answer key and scoring procedures. The latter requires judgment by content area experts. It is possible to have a positive item discrimination and a positive distractor-total correlation. Such an occurrence would indicate that an item does not have a clearly correct answer or that the response options have been stated in an ambiguous manner. The possibility of having a positive distractor-total correlation and a positive item discrimination is the main reason for conducting a distractor analysis.

Norm-referenced tests require individual differences on the measured construct. **If everyone were to obtain the same score on the test, this type of interpretation fails.** There is no way to rank order examinees when they have the same score. A norm-referenced test should be designed to **maximize score variance.** Item difficulty and discrimination are commonly used in conjunction when selecting items for a norm-referenced test. Item discrimination values should be between 0.3 and 0.7, and binary items should have a difficulty value near 0.5. If guessing is a factor, then the ideal difficulty value is slightly more than halfway between chance and 1 (Cronbach & Warrington, 1952; Lord, 1952). For example, a multiple-choice item with four options has an ideal difficulty value that is slightly more than [0.25 − (1 − 0.25)/2] = 0.625. T**hese rules of thumb are based on the idea that these values maximize item variance and subsequently increase score reliability.** Of course, every item will not likely have an item difficulty that is exactly the ideal value, but items with a difficulty within 0.2 points would be acceptable. For example, for a multiple-choice question in which guessing is not a factor, difficulty values between 0.3 and 0.7 would be acceptably close to the ideal value of 0.5. This range of values will maximize information that the test provides about score differences . No comparable rules of thumb exist for polytomous items, but by extension, it is best to select polytomous items that have moderate discrimination values and large item variances

In a criterion-referenced test, scores are not interpreted relative to the group of examinees. They are interpreted with respect to the content domain and often an absolute standard of performance referred to as a passing score. This type of test is common in education as interest lies in determining which students pass the test and which ones do not. That is, the interest is on mastery of the tested content. For example, a student who earns a score of 85% correct is considered to have mastered about 85% of the content domain. If the passing score was set at 80%, then this student is also considered to have passed the test. It is important to note that every student taking the test can get a score of 85% and pass the test. Since there is no comparison among examinees, it is perfectly acceptable for everyone to achieve the same score. Criterion-referenced tests emphasize the content domain and the passing score; they are **not designed to maximize score variability.** Consequently, item selection guidelines for a criterion-referenced test differ from a norm-referenced test. In criterion-referenced testing with a passing score, it is **not appropriate to choose items simply for the purpose of maximizing variance** (Crocker & Algina, 1986, p. 329). **Item selection should account for the passing score** and **aim to select items that maximize decision consistency (i.e. consistency in the pass/ fail decisions).** A variety of alternative item statistics are available for describing the **relationship between an item and the passing score**, such as the **agreement index,**

**B-index, and phi-coefficient.** A recent study evaluated the use of these statistics for item selection and compared the results to item selection based on more traditional statistics such as item discrimination. The results showed that the phi-coefficient tended to select easy items and was more dependent on the cut score than item discrimination . Moreover, using the phi-coefficient and B-index to select items for a test only improved decision consistency by 1% over item discrimination . **Taken together, these results suggest that item discrimination is also a useful index for criterion-referenced tests and choosing items with discrimination values between 0.3 and 0.7 will lead to high levels of decision consistency**. With respect to item difficulty and criterion-referenced tests, an argument can be made for **choosing items that span a wide range of difficultie**s. **Examinees of all ability levels will then be able to answer some items, but only accomplished students will be able to answer most of them. As with norm-referenced tests, item difficulty is not the primary consideration when selecting items for a test. Rather, statistics that describe the relationship between the item and total score (or passing score) should have greater weight in the decision to keep or eliminate an item from the test. After deciding which items to keep and which ones to eliminate from the test, an item analysis should be conducted on the final selection of test items.**

# Reliability

Reliability is a massive topic and it depends on our theory and measurement model for which reliability we might use. Finding reliability for the test would depend on which measurement model we are using. There are multiple methods for finding reliability for different measurement models. There are some types of reliability which we can find for all models. I will be assuming we are using primarily CTT for this test. If we were using IRT I might have used each error associated with each person(theta) or a simple mean of errors across the person continuum also called the **separation index**. If i was using structural equation modelling i would be using **Composite Reliability.** Reliability will be calculated with different methods if this were a criterion test.

If errors might be suspected due to time: - We could have found the test–retest reliability which requires participants to be tested and then retested.

If errors are suspected due to Observers: - Since this is an objective measure, we can rule out any error due to observers. If error was due to the observer though, we could have used two methods for this. If this was a unidimensional measure, we could have used **Multi-faceted Rasch models** to account for observer error. We can also use Interrater Reliability. To obtain the interobserver reliability coefficient, researchers calculate a correlation coefficient between the different raters' judgments of the same behaviors and then square that correlation. However, the interrater reliability has serious issues. They can demonstrate only consistency among the rank orders of candidates. They do not tell us anything about the severity or leniency differences between judges. A Multi-faceted Rasch models would be a much better choice in such a case

A **split-half reliability** estimate requires only a single test administration. After examinees complete the exam, it is divided into two halves, and the half-test scores are correlated. Given that correlation coefficients are affected by a restriction in range, it is necessary to adjust the half-test correlation to account for the full length of the test. This adjustment is achieved with the Spearman-Brown formula (Brown, 1910; Spearman, 1910), which is given by SB = 2r/(1+r), where r is the half-test correlation. For a half-test correlation of 0.6, the Spearman-Brown estimate is 0.75.

Internal consistency estimates of reliability are an alternative to split-half methods. They use inter-item covariances to produce an estimate of reliability. They are referred to as internal consistency estimates because they reflect how well items hang together. A common misconception about internal consistency is that these estimates reflect the extent to which a measure is unidimensional. That is, people often confuse a high internal consistency estimate as an indication that the test measures only one thing. Tests that measure multiple dimensions can also produce high internal consistency estimates. Therefore, the proper interpretation is to consider them a measure of the **similarity** among items. If there is a high degree of **similarity**, then items will produce consistent scores.

If errors are suspected to be due to participants: Internal consistency essentially measures how well do items go along with each other.

The earliest internal consistency methods date back to work by Kuder and Richardson who developed a number of methods including the KR20 and KR21. These two methods are limited to binary items, but the KR21 makes the additional assumption that items are equally difficult. Guttman expanded on their work and developed six lower bounds to reliability. His third lower bound (L3) is better known as coefficient alpha, and it generalizes the KR20 to binary and polytomous items. Stated differently, coefficient alpha and the KR20 produce the same result when items are binary.

Guttman's third lower bound is a simplification of his second lower bound (L2). However, his second lower bound is actually a better lower bound to reliability. L2 will always be equal to or greater than coefficient alpha (L3).

Cronbach demonstrated that coefficient alpha can be high even when a test measures multiple dimensions. Thus, he warned that it should never be interpreted as an indication that a test measures only one thing.

Methods for estimating reliability make different assumptions about the relationships among scores, and these affect the interpretation of an estimate. Many estimates are based on the assumption that scores are parallel. Lord and Novick described this assumption as having scores that “measure exactly the same thing in the same scale and, in a sense, measure it equally well for all persons” (1968, p. 48). **Alternate forms** and the **Spearman-Brown formula** are two methods based on this assumption. However, it is a very strict assumption that is unlikely to be met in practice. **Tau-equivalence** relaxes the assumption of parallel measures by allowing scores to have different error variances. Essential tau-equivalence further relaxes the assumption by allowing for unequal error variances and different mean scores. The least strict assumption is that scores are congeneric. **Congeneric measures allow for different score means, error variances, and relationships with the latent trait**. It is arguably the most tenable assumption in practice. Coefficient alpha estimates reliability when scores are at least essentially tau-equivalent, but w**hen scores are congeneric, coefficient alpha is only a lower bound to reliability.** Guttman’s coefficients were all developed under a relaxed set of assumptions and should always be considered to be lower bounds.

There are other measures of internal consistency that were developed under the assumption of congeneric or classically congeneric measures. Two of these methods are provided in jMetrik. Feldt and Gilmer developed a measure of internal consistency using an assumption of classically congeneric measures. It is an assumption like congeneric measures but with the added restriction that error variances be proportional to effective test length. The Feldt-Gilmer coefficient is difficult to compute and can only be computed when tests have more than two parts (e.g. more than two items). An easier method of computing a reliability estimate for classically congeneric measures is the **Feldt Brennan coefficient.**

The downside of the SEM is that it depends on the scale of measurement. We cannot use it to compare the quality of measures that use different scales. We cannot even use it to compare a measure to a shortened version of the same measure because shortening a test changes the observed score scale. We must use a reliability estimate to compare measures that are on different scale because the reliability coefficient is scale free.

The reliability coefficient is affected by several factors. As evident in the equation for the reliability coefficient, reliability will be larger for heterogeneous populations , and it will be smaller for homogenous populations. This factor is a reason for **reporting a reliability estimate for different subgroups of examinees .** For example, males may be more heterogeneous than female examinees. As a result, reliability will be higher for males than it will be for females. Reporting an estimate for each group is necessary to ensure that reliability is at an acceptable level for all examinees. As second factor that affects reliability is **test length. All other things being equal, longer tests are more reliable than shorter tests.** An important caveat to note is that reliability will not increase by adding any arbitrary item to the test. The new items must be like the ones already on the test. T**wo other factors that affect reliability are related to the underlying assumptions of classical test theory: dimensionality and uncorrelated errors**. Unidimensionality requires that a test measures one and only one construct. If a test measures

multiple constructs (whether intentionally or not), reliability will be underestimated. That is, reliability will be a lower bound when multiple dimensions affect test scores. **The assumption of unrelated error scores means that our test items are independent.** An examinee's response to one question does not affect her response to another one. This assumption tends to be **violated** in practical testing situations. It may be violated under conditions of **test speededness,** which occurs when a test's time limit is too strict and examinees change their response behavior near the end of a test. It may also be violated when a test contains groups of related items such as a group of items that ask questions about a common reading passage or a group of math items that pertain to a shared graph or figure. These groups of related items are called **testlets** , and they are like a test within a test. The effect of correlated errors on reliability estimation is less predictable than the effect of multidimensionality. In some cases, correlated errors will overestimate reliability, and in other cases, they will under estimate it.

**Reliability for criterion tests**

In a criterion-referenced framework, the focus of reliability concerns whether or not we would make the same pass/fail decision if the test were replicated. Returning to the concept of repeating the measurement procedure over and over again, a consistent decision is made whenever an examinee fails every time the test is repeated. A consistent decision is also made whenever an examinee passes every repetition of the test. Errors are made when an examinee fails some repetitions of the test but passes others. These errors represent inconsistent decisions. As described earlier, repeating the test many times does not actually occur, **but to estimate decision consistency, we need at least two replications of the measurement procedure**. As done when computing alternate forms reliability, two very similar tests can be constructed and given to examinees. A raw agreement index can be computed by summing the number of examinees who pass both tests or fail both tests and dividing this number by the total number of examinees. **This statistic would be 0 when everyone fails one test but passes the other (or vice versa). It would be one when all of the examinees pass both tests or fail both tests.** A limitation of the raw agreement index is that it does not account for chance levels of agreement. That is, **we expect to have a certain amount of agreement that is due to chance and not the quality of a test.** Cohen (1960) developed an alternative statistic that adjusts for chance levels of agreement. His kappa statistic is typically smaller than the raw agreement index, but it has a different interpretation. **For example, a kappa statistic of 0.3 is interpreted as a 30% improvement beyond chance. Compared to the raw agreement index, 0.3 seems small, but a 30% improvement beyond chance is a substantial gain**. **Raw agreement and kappa statistics** are rarely used in practical testing situations because they **require examinees to take two tests.** Thankfully, researchers have created methods for estimating these statistics from a single test administration. In particular, Huynh (1976a, 1976b) developed methods based on the assumption that scores from one test follow a beta-binomial distribution and that scores from a second, hypothetical test follow the same beta-binomial distribution. Consequently, scores from both tests follow a bivariate beta-binomial distribution. Estimates of raw agreement and kappa may then be obtained from the **bivariate beta-binomial distribution.** Interpretation of **Huynh's raw agreement** and **kappa statistics** are the same as described earlier. The only difference is that his methods involve a hypothetical replication of the

test. Huynh's raw agreement and kappa statistics use the beta-binomial and bivariate beta-binomial distributions.

Caption Original

TEST LEVEL STATISTICS

| | |
|---|---|
| Number of Items | 6 |
| Number of Examinees | 291 |
| Min | 0.0000 |
| Max | 12.0000 |
| Mean | 9.6838 |
| Median | 11.0000 |
| Standard Deviation | 2.5799 |
| Interquartile Range | 4.0000 |
| Skewness | -1.2789 |
| Kurtosis | 1.2953 |
| KR21 | 2.2719 |

RELIABILITY ANALYSIS

| Method | Estimate | 95% Conf. Int. | SEM |
|---|---|---|---|
| Guttman's L2 | 0.8267 | (0.7939, 0.8558) | 1.0760 |
| Coefficient Alpha | 0.8215 | (0.7878, 0.8515) | 1.0918 |
| Feldt-Gilmer | 0.8259 | (0.7930, 0.8552) | 1.0783 |
| Feldt-Brennan | 0.8258 | (0.7929, 0.8551) | 1.0786 |
| Raju's Beta | 0.8215 | (0.7878, 0.8515) | 1.0918 |