Assessment
Systems
Corporation
www.assess.com

Classical Item and Test Analysis with CITAS

White Paper

Nathan A. Thompson, Ph.D.
Vice President, ASC
Adjunct Faculty, University of Cincinnati

Contact Information

Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, Minnesota 55114
Voice: (651) 647-9220
Fax: (651) 647-0412
E-Mail: solutions@assess.com
www.assess.com

Table of Contents

What is CITAS?

The Classical Item and Test Analysis Spreadsheet (CITAS) is a simple tool to statistically analyze small-scale assessments, available for free download here. CITAS is a Microsoft Excel® spreadsheet with all necessary calculations programmed in as formulas, which means that all the user is required to do it type or paste in the student responses and the correct answers, or keys. CITAS will then score all students with number-correct (NC) scoring, as well as populate important statistics. Statistics include both test-level statistics such as reliability, and item-level statistics such as difficulty (*P*) and discrimination (r_{pbis}). CITAS statistics are listed in Table 1.

Table 1: Statistics in CITAS output

Test-level statistics	***Item statistics***
Number of examinees	*P*
Number of items	r_{pbis}
NC score mean	Number correct
NC score standard deviation	Number incorrect
NC score variance	Mean score correct
Minimum NC score	Mean score incorrect
Maximum NC score	Response frequencies
KR-20 (alpha) reliability	Response mean scores
Standard error of measurement (SEM)	
Mean *P*	
Minimum *P*	
Maximum *P*	
Mean r_{pbis}	
Minimum r_{pbis}	
Maximum r_{pbis}	

This paper will begin by defining the concepts and statistics used in classical item and test analysis, and then present how the CITAS spreadsheet provides the relevant information. The purpose of CITAS is to provide an option for quantitative analysis of testing data that is as straightforward as possible.

Why is it called "classical?" This is to differentiate this type of analysis from the modern test analysis approach called *item response theory* (IRT). IRT is much more powerful, but only works with sample sizes numbering in the hundreds or larger. This makes it extremely important in large-scale testing, but completely inappropriate for classroom-sized samples.

Classical analysis at the test level

Classical test analysis is based primarily on the NC scores. CITAS calculates descriptive statistics of the NC scores, as well as two important indices from classical test theory: KR-20 (α) reliability, and the standard error of measurement (SEM).

Reliability is a classical concept that seeks to quantify the consistency or repeatability of measurement. If a test is producing consistent scores, then we say it is reliable. As to whether the scores actually mean what is intended, is part of a larger and more difficult issue called *validity*. What is meant by consistency? Let us assume that a student has a *true score* of 44 items out of 50. If they took the test multiple times (assuming that we wiped their memory of the test), they might get a 45, or a 43, etc. This is consistent. If they were to score a 34, then a 47, then a 39, it would be an unreliable test.

There are several approaches to indexing reliability, the most common of which is *internal consistency* using the Kuder-Richardson Formula 20 (KR-20) index. Another index called *coefficient α* is also common, but is nothing more than a generalization of KR-20 to *polytomous* (rating scale or partial credit) data.

KR-20 ranges in theory from 0.0 to 1.0, with 0.0 indicating random number generation and 1.0 indicating perfectly consistent measurement. In rare cases, it can even dip below 0.0. Therefore, a higher number is regarded as better. However, KR-20 is partially a function of test length, and tends to be higher when the test has more items. So for shorter tests like 20 or 50 items, it is unrealistic to expect KR-20 values near 1.0. In some cases, a value of 0.7 might be sufficient.

KR-20 is also important because it is used in the calculation of the SEM. The SEM takes the concept of measurement consistency and applies it to student scores. If we take plus or minus 2 SEMs around a student's observed score, that gives us a range we are 95% confident contains their true score. If this number is very small, this means that we have an accurate estimate of the true score. If it is large, we do not have an accurate estimate. Suppose the SEM is 5.0. Then the interval is plus or minus 10. For a student with a test score of 39, this means we expect their true score to be anywhere from 29 to 49: hardly an accurate test!

Classical test analysis with CITAS

CITAS provides both the KR-20 and SEM, as well as simple descriptive statistics of the student scores. Table 1 explains the summary statistics found on the "Output" tab

Table 2: Summary test-level statistics in CITAS output

Summary Statistic	***Definition***
Test-level	
Examinees:	Number of students
Items:	Number of items
Mean:	Average NC test score
SD:	Standard deviation of NC scores
Variance:	Variance of NC scores
Min:	Lowest score
Max:	Highest score
KR-20:	Reliability of measurement
SEM:	Standard error of measurement

Example output is shown in Table 3. This test has only 20 items, which were answered by 50 students. The test was quite easy, with an average of 18.24 out of 20. There was a moderate spread of scores, with an of SD = 2.24 and a range of 11 to 20. The test was very reliable given its short length, having a KR-20 of 0.73 and an SEM of only 1.17.

Table 3: Summary test-level statistics in CITAS output

Statistic	Value
Examinees:	50
Items:	20
Mean:	18.24
SD:	2.24
Variance:	5.00
Min:	11
Max:	20
KR-20:	0.73
SEM:	1.17

Classical analysis at the item level

The statistics presented so far only provide information at the level of the entire test. Classical test theory also has statistics for evaluating individual items from a quantitative perspective. The goal of item analysis is to use detailed statistics to determine possible flaws in the item. This can be something as specific as identifying a bad distractor because it pulled a few high-ability examinees, or something as general as "this item is harder than I like for my students."

In large-scale or high-stakes testing, item analysis is typically performed before the test goes "live" to ensure only that quality items are used. Often, it is done after pretesting the items on some small set of the population, perhaps by inserting them as unscored into the test the year before they are to be used as scored items. In cases where pretesting is not feasible, item analysis (as well as equating and standard-setting) can take place after the live administration. However, this requires that scores be reported much later, in some cases weeks or months later. This is of course unacceptable for classroom assessment, but CITAS allows you to evaluate items immediately after test delivery.

Item analysis is important because it is analogous to quality control of parts used in the assembly of a final manufacturing product. Nobody wants bad tires or bad brake pads in their car. Releasing tests with bad or untested items is like releasing cars off the assembly line with bad or untested brake pads. Both prevent plenty of opportunities for litigation.

In classical item analysis, there are two concepts we are interested in assessing: item difficulty and item discrimination. *Item difficulty* is a simple concept in classical test theory; it simply refers to the proportion of students that correctly answered an item. This is called the *P*-value. Yes, I know, we already use "*p*-value" as the term for statistical significance, but it's too late to change a 100-year-old theory. If 95% correctly answered it, the item is quite easy. If 30% correctly answer, then the item is quite difficult, especially when you consider that a four-option multiple choice item presents a 25% chance of guessing the correct answer! For this reason, items with $P < 0.50$ are generally considered to be quite difficult, while we typically see them more in the 0.70 to 0.80 range. However, specific tests might produce different ranges of statistics, requiring you to shift the paradigm somewhat.

Just what is too difficult or too easy? That is a judgment call that you have to make while taking into account the content of the item, the purpose of the test, and the sample of students. A test that is designed to be extremely difficult might have an average *P* of 0.60. Conversely, a test that is given to a group of extremely high ability examinees can be expected to have an average *P* of 0.90. Regardless of the average *P*, it is often preferable to have items with a range of difficulty. If you have no items with $P < 0.70$, it means that all the items were fairly easy, and there was not a single item on the test to "separate the men from the boys." This might be acceptable if the purpose of the test is just to assess entry level knowledge to a topic, but if the purpose of the test is to identify the top students, a test with all item *P* values above 0.70 would not do a good job.

Item discrimination refers to the power of the item to differentiate between examinees with high and low levels of knowledge or ability. But of course we never know the true score of any examinee. The best estimate we have is the total NC score on the test. Therefore, item discrimination is typically defined as the correlation between item scores (scored 0/1) and total test scores, called the *item-total correlation*. The equation used to calculate this is called the *point-biserial correlation*, or r_{pbis}, though some researchers prefer to use its cousin called the *biserial correlation*. This provides an index of whether students who get the item correct are scoring highly, which is the hallmark of a good item.

The item-total correlation has the advantage that it is interpreted as any other correlation, which many people are used to working with. An r_{pbis} of 0.0 indicates that there is no correlation, which means that there is no relationship between the item and the total scores. This means that the item is providing no information, and item responses are essentially random with respect to total scores. But as r_{pbis} increases, it indicates a stronger relationship between the item and total score. A value of 0.20 means a decent item, and highly discriminating items will have values in the 0.50 or 0.60 range.

On the other hand, a negative r_{pbis} is very bad news. This means that there is an inverse relationship, namely that low-scoring students performed better on the item than high-scoring students did. This typically indicates one of three things:

1. A key error;
2. A very attractive distractor;
3. This item is so easy/hard that there are few examinees on one side of the fence, making it difficult to correlate anything.

All three things are issues with the item that need to be addressed.

For completeness' sake I must mention another statistic, sometimes called the *classical difficulty index* or the *top-bottom index*. This was developed before the r_{pbis} but is occasionally still used. It is based on the same concept that we want high-ability students to get the item correct more often than low-ability students. So we divide the sample in half, and find the proportion of the top half correctly answering and subtract the proportion of the bottom half correctly answering (sometimes done with the top and bottom 27%). Like the r_{pbis}, a positive value indicates a better quality item.

A very important thing to note about the item statistics: like all statistics, their stability depends on sample size. In general, we need 20 or 30 people to get marginally useful statistics, and they start becoming statistically stable near 50 examinees. Therefore, while CITAS results with 20 students will provide some helpful information, do not consider the item statistics to be perfectly stable.

Classical item analysis with CITAS

CITAS presents the *P* and r_{pbis} for each item as well as some supplemental statistics based on the correct/incorrect dichotomy. The first is the number of correct and incorrect responses. This is obviously a repackaging of the *P* value, but provides an alternative method of looking at difficulty if you prefer to use it. Additionally, CITAS presents the mean scores for students who got the item correct and incorrect. If the item is discriminating well, the mean score will be higher for the "correct" students. Similarly, this is a repackaging of the r_{pbis}, but provides an alternative method of evaluating item discrimination.

Table 4 presents example results from a test of 20 items. Let us go through the results for the first five items and interpret the statistics.

Item 1 is a fairly easy item with a *P* of 0.92, and has a minimal but still positive r_{pbis} at 0.13. This positive discrimination is reflected in the mean scores; the average score for examinees responding correctly is about 1 point higher than examinees responding incorrectly. This is a fairly small difference, but indicates that the item is still acceptable.

Item 2 is similar to Item 1, but is slightly more discriminating, with an r_{pbis} of 0.27 and a point difference between the groups of more than 2 points. This item is a solid item for the test even though it is relatively easy.

Item 3 presents an ideal item from the classical perspective. It is more difficult than the first two items, but still not all that difficult in an absolute sense with 80% of the examinees responding correctly. More importantly, it has a very strong discrimination; the r_{pbis} is 0.62 and the point difference is nearly 3.5 points. Items like this are very powerful, as we can achieve decent test reliability with only a few items. Note that this example test has a reliability of 0.73 even though it is only 20 items. This is because the average r_{pbis} is a notable 0.39.

Items 4 and 5 present an example of what happens when an item is too easy. Because there are only 1 or 2 examinees that responded incorrectly, there is very little differentiating power. We can see that item 5 had only one person respond incorrectly, and they had a score of 19, which led to a negative r_{pbis}.

The remainder of the items in this test presents examples similar to items 2 and 3. This test is composed of items that are quite easy with an average *P* of 0.91 and medium to strong discrimination statistics. This is indicative of a good but easy classroom test, with the exception of items 4 and 5.

Table 4: CITAS item statistics

Item	P	Rpbis	Number correct	Number incorrect	Mean score correct	Mean score incorrect
1	0.92	0.13	46	4	18.33	17.25
2	0.92	0.27	46	4	18.41	16.25
3	0.80	0.62	40	10	18.93	15.50
4	0.96	0.07	48	2	18.27	17.50
5	0.98	-0.05	49	1	18.22	19.00
6	0.88	0.43	44	6	18.59	15.67
7	0.98	0.34	49	1	18.35	13.00
8	0.92	0.27	46	4	18.41	16.25
9	0.92	0.40	46	4	18.50	15.25
10	0.90	0.37	45	5	18.51	15.80
11	0.92	0.50	46	4	18.57	14.50
12	0.86	0.54	43	7	18.72	15.29
13	0.96	0.53	48	2	18.48	12.50
14	0.92	0.60	46	4	18.63	13.75
15	0.94	0.48	47	3	18.51	14.00
16	0.84	0.39	42	8	18.62	16.25
17	0.96	0.53	48	2	18.48	12.50
18	0.84	0.49	42	8	18.71	15.75
19	0.86	0.38	43	7	18.58	16.14
20	0.96	0.58	48	2	18.50	12.00

Distractor analysis with CITAS

In addition to evaluating items as a whole, statistics can be used to evaluate individual *options* of items. The statistics for the correct option serve as the statistics for the item as a whole, because 90% of the students answered the correct answer of "A" then 0.90 is both the proportion of students who answered "A" and the proportion of students who answered correctly. But what makes option statistics useful is the evaluation of the incorrect options, known as *distractors*. This provides even greater detail about the performance of the item, as we will see.

CITAS provides a method to evaluate distractors by presenting the average scores for examinees with a given response. If the item is performing well, examinees that respond correctly will have the highest average score. Examinees responding incorrectly will have a lower average score. When examining individual options, the option that is the most incorrect should have the lowest average score. For example, if "A" is correct, "B" and "C" are incorrect, and "D" is not even close, then we would expect student who selected "A" to have high scores, and students who selected "D" to have low scores.

The final tab of CITAS presents statistics for distractor analysis, as seen in Table 5. We saw in Table 4 that 46 examinees responded correctly to the first item, while 4 responded incorrectly. Table 5 shows us that of those 4 incorrect responses, 3 chose "B" and 1 chose "D." In this case, not a single person in our sample responded with a "C." This does not make it a bad item or even a bad distractor, but could indicate that future items written on the same objective might want to have a different thought process in the development of distractors.

Table 5: CITAS distractor analysis for example form 1

		Frequencies				***Mean Score***			
Item	**Key**	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
1	A	46	3	0	1	18.33	17.00		18.00
2	D	1	1	2	46	11.00	18.00	18.00	18.41
3	C	3	5	40	2	18.00	14.20	18.93	15.00
4	D	1	0	1	48	17.00		18.00	18.27
5	B	1	49	0	0	19.00	18.22		
6	D	1	4	1	44	17.00	14.75	18.00	18.59
7	A	49	0	1	0	18.35		13.00	
8	C	1	1	46	2	17.00	19.00	18.41	14.50
9	D	2	1	1	46	15.00	14.00	17.00	18.50
10	B	1	45	4	0	17.00	18.51	15.50	
11	A	46	2	2	0	18.57	16.50	12.50	
12	A	43	3	4	0	18.72	18.33	13.00	
13	D	1	0	1	48	13.00		12.00	18.48
14	C	0	2	46	2		15.00	18.63	12.50
15	C	0	2	47	1		15.50	18.51	11.00
16	D	3	4	1	42	16.00	17.00	14.00	18.62
17	A	48	1	0	1	18.48	13.00		12.00
18	B	2	42	4	2	15.50	18.71	15.75	16.00
19	B	1	43	5	1	17.00	18.58	15.80	17.00
20	B	1	48	1	0	11.00	18.50	13.00	

Similarly, Table 5 shows that the average score for the correctly responding examinees was 18.33. We also saw this in Table 4, but Table 5 provides the average scores for the "B" and "D" responding examinees separately to evaluate the individual distractors.

With so few examinees selecting the incorrect options, the averages for those options have little meaning. Item 3 presents a better example for this type of analysis. We still see that the average score for those selecting the correct option is 18.93. But we are able to see just what types of examinees are selecting the incorrect options. Those selecting "A" are of fairly high ability, with an average score of 18.00, whereas those selecting "B" and "D" have the much lower averages of 14.20 and 15.00. We can then go back to the item and ask ourselves if "A" is possibly too attractive of an option for high ability examinees, and also examine if "B" and "D" are that strongly incorrect.

But all of the items in this example set are good items, as we can see by the strong r_{pbis} values with the exception of Item 5 – which only had a low r_{pbis} because it was too easy. So what would a bad item look like? It would have a negative r_{pbis} but still a moderate difficulty. This would mean that more high ability examinees were selecting one of the distractors than were selecting the correct response. This would possibly indicate that the distractor in question is arguably correct or that the stem of the question is quite unclear. There could be various issues along these lines, and items with bad statistics need to be reviewed closely, one by one.

Summary statistics of item statistics

CITAS provides one additional level of information: summary statistics of the individual item statistics. These are presented in the lower right of the "Output" tab. Example results are show in Table 6. We again see that this is an easy test; the average *P* value was only 0.91, meaning that the average score was 91% correct. The most difficult item had a *P* of 0.80 and the easiest item had a *P* of 0.98.

The discrimination power of this test was quite good even though it was of little difficulty (items that are too easy will limit the r_{pbis} because there is nothing to correlate). The average r_{pbis} was 0.39, and the lowest was item 5 with a value of -0.05. The most discriminating item was item 3 with 0.62.

Table 6: Summary statistics in CITAS output

Summary Statistic	*Value*
Mean P:	0.91
Min P:	0.80
Max P:	0.98
Mean Rpbis:	0.39
Min Rpbis:	-0.05
Max Rpbis:	0.62

Summary

Item analysis is a vital step in the test development cycle, as all tests are composed of items and good items are necessary for a good test. Classical test theory provides some methods for evaluating items based on simple statistics like proportions, correlations, and averages. However, this does not mean item evaluation is easy. I've presented some guidelines and examples, but it really comes down to going through the statistical output and a copy of the test with an eye for detail. While psychometricians and software can always give you the output with some explanation, it is only the item writer, instructor, or other content expert that can adequately evaluate the items because it requires a deep understanding of test content.

Although CITAS is quite efficient for classical analysis of small-scale assessments and teaching of classical psychometric methods, it is not designed large-scale use. That role is filled by two other programs, *Lertap 5* and *Iteman 4*. *Lertap 5* is an Excel-based system designed for comprehensive classical analysis; you can learn more at its website here. *Iteman 4* is also designed to produce a comprehensive classical analysis, but in the form of a formal report ready for immediate delivery to content experts; see an example report at its website here.

Further reading

Downing, S.M., & Haladyna, T.M. (Eds.) (2006). *Handbook of test development*. Philadelphia: Taylor & Francis. (website)

Furr, R.M., & Bacharach, V.R. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage. (website)

Shultz, K.S., & Whiney, D.J. (2005). *Measurement theory in action*. Thousand Oaks, CA: Sage. (website)