EAC's of Assessment Statistics
Table of Contents

KR-20 / Cronbach Alpha for Overall Reliability (Exams and Rubrics)

The **KR-20**, or Kuder-Richardson Formula, and the **Cronbach Alpha** measure overall test and rubric reliability. These measures let you know whether the exam or the rubric as a whole discriminated among students who mastered the subject matter and those who did not.

The KR-20/Cronbach Alpha value is shown on the **Summary Statistics** menu for a Rubric or Exam in EAC Visual Data.

The KR-20 / Cronbach Alpha value generally ranges between **0.0** and **+1.0**, but it can fall below 0.0 with smaller sample sizes.

The closer the value is to +1.0 the more reliable an exam or rubric is considered because its questions or criterion rows do a good job of consistently discriminating among higher and lower performing students.

For example, a KR-20 of 0.0 means that the exam questions or the rubric row criteria didn't discriminate at all. Imagine a test where all 20 students answered all 40 questions correctly. This test doesn't discriminate among any of the students, so its KR-20 of 0.0 makes perfect sense. Or imagine a 4-row rubric with a .90 KR-20. This suggests that when the students were scored on the rows, low-performing students were scored at the lower levels of achievement and higher performing at the higher level. The rubric's descriptors for each level of achievement in each of the rows were effective for the evaluator to determine the appropriate level.

EAC suggestion: The interpretation of the KR-20 depends on the purpose of the test and the rubric.

- **For Exams**: Most high stakes exams are intended to distinguish those students who have mastered the material from those who have not. For these, shoot for a KR-20 of **+0.50 or higher**. A KR-20 of less than +0.30 is considered poor no matter the sample size. If the purpose of the exam is to ensure that ALL students have mastered essential skills or concepts or the test is a "confidence builder" with intentionally easy questions, look for a KR- 20 close to 0.00.
- **For Rubrics**: If the rubric has a score of +0.50 or higher, then there is good internal consistency among the descriptors in the levels of achievement and the criteria / rows in the rubric. The evaluators have been able to effectively differentiate among the levels when scoring students.

Cronbach Alpha with Deletion (Exams and Rubrics)

Cronbach Alpha with Deletion helps assess the reliability of individual test questions and individual rubric rows.

For Test Questions: The Cronbach Alpha with Deletion ("Cronbach Del") value is presented on the **ItemAnalysis** menu.

How does Cronbach Alpha with Deletion work? It asks whether the exam as a whole is more reliable if you simply **delete** the question under review. The Cronbach Alpha with Deletion re-runs the exam's KR-20 without the question under review. If the exam as a whole is more reliable without it, there'sprobably something wrong with that question.

The Cronbach Alpha with Deletion generally ranges between **0.0** and **+1.0**, but it can fall below 0.0 with smaller sample sizes.

More important than its range is how the Cronbach Alpha with Deletion compares to the exam's KR-20. If a question's Cronbach Alpha with Deletion is **greater than** the exam's KR-20, it means the exam as a whole is more reliable without it. For example, look at Question No. 1 in the picture below. If the exam's KR-20 is 0.463, then we know Question No. 1 is "suspect" because its Cronbach Alpha with Deletion of 0.52 is greater than 0.463.

Biology 1 Pre-Test (1/25/2019 - 2/8/2019)

Summary Statistics

Description	Value	Description	Value
Scored Responses	20.00	Highest Score	4.00
Scorable Questions	4.00	Lowest Score	1.00
Possible Item Scores	80.00	Mean Score	2.55
Actual Item Scores	80.00	Median Score	2.00
Std Dev	0.81	KR(20) / Cronbach Alpha	0.463

Item Analysis

DI 27 Jump To 1

No.	Title	Question	P-Value	Point Biserial	Cronbach	DI (27)
1		What is 5 + 5?	1.00	0.00	0.52	0.00
2		What is 30 - 15?	0.95	0.19	0.46	0.20
3		Select all sums greater tha…	0.45	0.48	0.12	1.00
4		Select all differences less …	0.15	0.43	0.20	0.60

EAC suggestion: Look out for questions with a Cronbach Alpha with Deletion greater than the exam's KR-20. These questions decrease overall test reliability and should be considered suspect.

For Rubrics: The Cronbach Alpha with Deletion helps assess the reliability of each rubric row. Like a test item Cronbach Alpha with Deletion, it re-runs the KR-20 but does so without the rubric row under review.

ISLO Writing Assessment (1/25/2019 - 10/6/2019) STANDARD SUMMARY FULL CURR

Summary Statistics

Description	Value	Description	Value
Scored Responses	67.00	Highest Score	6.00
Rows	6.00	Lowest Score	2.25
Possible Item Scores	402.00	Mean Score	4.19
Actual Item Scores	392.00	Median Score	4.25
Std Dev	0.91	KR(20) / Cronbach Alpha	0.79

Row Analysis   Jump To 1

No.	Row	Average	Point Biserial	Cronbach Del
1	Focus	0.78	0.73	0.72
2	Organization	0.72	0.66	0.74
3	Support	0.66	0.71	0.74
4	Critical Thinking	0.78	0.38	0.79

EAC suggestion: Look out for rubric rows with a Cronbach Alpha with Deletion greater than the rubric's KR-20. These rows decrease the overall reliability of the rubric.

Point Biserial Correlation (Exams and Rubrics)
The **point biserial correlation** measures item reliability.

For Exams, the point biserial correlation compares student scores on one particular question with their scores on the test as a whole. The Point Biserial Correlation can be found on the **Item Analysis** menu.

The driving assumption behind **Point Biserial** is simple: Students who score well on the test as a whole should on average score well on the question under review. Students who struggle on the test as a whole should on average struggle on the question under review. If a question deviates from this assumption (aka, a "suspect" question), the point biserial correlation lets us know.

The point biserial correlation ranges from a low of **-1.0** to a high of **+1.0**. The closer the point biserial correlation is to +1.0 the more reliable the question is considered because it discriminates well among students who mastered the test material and those who did not.

A negative point biserial correlation means that students who performed well on the test as a whole tended to miss the question under review and students who didn't perform as well on the test as a whole got it right. It's a red flag, and there are a number of possible things to check.

- Is the answer key correct?
- Is the question clearly worded?
- If it's multiple choice, are the choices too similar?

EAC suggestion: For high stakes exams intended to distinguish among students who mastered the material from those who did not, shoot for questions with point biserial correlations **greater than +0.30**. They're very good items. Questions with point biserial correlations less than +0.09 are considered poor. Questions with point biserial correlations between +0.09 and +0.30 are considered reasonably good.

For **Rubrics**, the point biserial correlation can be found on the **Row Analysis** menu.

ISLO Writing Assessment (1/25/2019 - 10/6/2019)

Row Analysis

No.	Row	Average	Point Biserial	Cronbach Del
1	Focus	0.78	0.73	0.72
2	Organization	0.72	0.66	0.74
3	Support	0.66	0.71	0.74
4	Critical Thinking	0.78	0.38	0.79
5	Conventions	0.64	0.71	0.72

The same principle applies for the analysis of individual rubric rows as it does for individual exam questions. With the range from a low of **-1.0** to a high of **+1.0**, the closer the point biserial correlation for a rubric row is to +1.0 the more reliable the row is in discriminating between levels of achievement for the students who performed well and those who did not.

p-Value (Exams Only)

In the branch of statistics dealing with test reliability, and unlike other branches of statistics, **p-Value** is a simple measure of question difficulty. P-Value is found on the **Item Analysis** menu.

Biology 1 Pre-Test (1/25/2019 - 2/8/2019)

Item Analysis   DI 27   Jump To 1

No.	Title	Question	P-Value	Point Biserial	Cronbach Del	DI (27)
1		What is 5 + 5?	1.00	0.00	0.52	0.00
2		What is 30 - 1...	0.95	0.19	0.46	0.20
3		Select all su...	0.45	0.48	0.12	1.00
4		Select all diff...	0.15	0.43	0.20	0.60

The p-Value ranges from a low of **0.0** to a high of **+1.0**.

The closer the p-Value is to 0.0 the more difficult the question. For example, a p-Value of 0.0 means that no student answered the question correctly and therefore it's a really hard question. If an item's p-Value is unexpectedly close to 0.0, be sure to check the answer key.

The closer the p-Value is to +1.0 the easier the question. For example, a p-Value of +1.0 means that every student answered the question correctly and therefore it's a really easy question.

EAC suggestion: On high stakes exams, shoot for p-Values between **+0.50 and +0.85** for most test questions. A p-Value less than +0.50 means the question may be too difficult or you should double check the answer key. A p-Value greater than +0.85 means the question may be too easy or most students have mastered that concept.

Distractor point biserial correlation (Exams Only)

The **distractor point biserial correlation (PtBis)** appears on the **Distractors** menu**.**
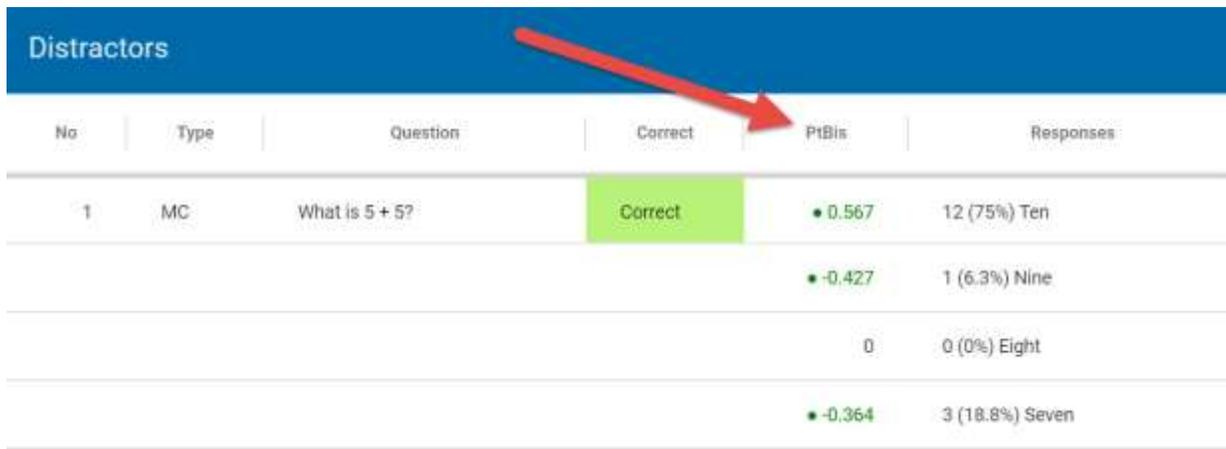
The **Distractors Point Biserial Correlation** digs deeper than the item statistics and measures the reliability of **each answer choice** presented to students. How? It correlates student scores on each answer choice with their scores on the test as a whole.

The driving assumption is simple: Students who score well on the test as a whole should on average select the correct answer choice for each question. Students who struggle on the test as a whole should on average select an incorrect answer choice for each question. If an answer choice deviates from this assumption, the distractor point biserial correlation lets us know.

The distractor point biserial correlation ranges from a low of **-1.0** to a high of **+1.0**. The closer a **correct** answer choice's distractor point biserial correlation is to **+1.0** the more reliable the answer choice is considered because it discriminates well among students who mastered the test material and those who did not. This answer choice "works well."

By the same token, the closer an **incorrect answer** choice's distractor point biserial correlation is to **-1.0** the more reliable the answer choice is considered because it discriminates well among students who did not master the test material and those who did.

EAC suggestion: Consider changing answer choices that aren't "working" as expected and also those that students don't select at all. If no student selected an answer choice, that answer choice isn't really a "distractor" after all.

NOTE: Distractor Point Biserial correlations can only really be generated for Multiple Choice test questions where more than 2 answers were selected by students. Thus, they are often blank.