Understanding Test Items – Handout

Ben O. Smith

February 2020
Updated September 2023

This handout provides the practitioner a working understanding of testing metrics often produced by exam and statistical packages. Some test metrics can help diagnose bad questions, some can help measure student ability, and some can measure student knowledge and/or learning.

Author Information:
Ben O. Smith, Associate Professor
bosmith@unomaha.edu
University of Nebraska - Omaha

P: https://bensresearch.com
D: https://cba.unomaha.edu/econ

Understanding test metrics from software programs is surprisingly difficult. Since 1904,[1] educators have been trying to find ever improving ways of measuring test quality, distractor quality, and therefore, knowledge or learning. Unfortunately, software packages often include outdated measures along with more modern measures. This is often confusing, especially given that measurement disturbance (error) is rarely provided.

In this document, we will quickly discuss test measurements that are the most likely to be provided by testing and statistical software. There are four basic types of test measures: (1) those based on Classical Test Theory, (2) those based on Item Response Theory, (3) those based on the Flow of Knowledge or Value-Added Learning, and (4) rubric-based knowledge measures.

[1] Ross E. Traub. Classical test theory in historical perspective. *Educational Measurement: Issues and Practices*, 16 (4):8–14, 1997. DOI: 10.1111/j.1745-3992.1997.tb00603.x

Classical Test Theory (CTT)

Classical Test Theory is based on a relatively basic idea: an exam score can be thought of as the sum of the true score and error.

$$\underbrace{S}_{\text{Score}} = \underbrace{T}_{\text{True Score}} + \underbrace{E}_{\text{Error}} \tag{1}$$

The true score is the students' true knowledge of the exam content. This is a latent variable and thus not directly observable. Error is anything that can distort the exam's measure of the students' true knowledge; E can be a positive or negative value. This can include confusing questions, question options that can be removed, or student guessing. Therefore, error is intrinsically 'bad' in this context as it adds noise to T. This basic insight results in one of the most common measures produced by testing software: *reliability*. Because of the additive nature of the model, variance is also additive. Therefore, variance can be described as:

$$\sigma_S^2 = \sigma_T^2 + \sigma_E^2 \tag{2}$$

If we want to measure the ratio of 'signal to noise,' we can describe it as:

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{3}$$

Fundamentally, CTT is trying to get at ρ. Unfortunately, there is no real way to know ρ as neither σ_T^2 or σ_E^2 are known. However, there are multiple estimates of ρ based on different assumptions. The two most important categories are based on either *test-retests* or *internal validity*. In the case of test-retest, the assessment is given to the same group of students twice (allowing sufficient time to pass between settings for the subjects to forget the questions themselves). Then the Pearson correlation coefficient (i.e. Pearson's r) is calculated between the two exams. The basic theory here is that T should be the same for both exams while E should be different.

Let's focus on measures of internal validity as possibly the most common measures generated by exam software. Probably, the most common measure is Cronbach's α.[2,3]

$$\hat{\alpha} = \frac{q}{q-1}\left(1 - \frac{\sum_i^q \sigma_{S_i}^2}{\sigma_S^2}\right) \tag{4}$$

Where q is the number of items (questions), $\sigma_{S_i}^2$ is the variance of item i and σ_S^2 is the variance of the overall test scores. Notably, this simply measures the degree to which this particular test item moves with the rest of the exam. In general, one would expect well-behaving items to produce a high score (above about 0.6). However, an $\hat{\alpha}$ can be too high as well. Given this is as measure of internal consistency, if one tests the same material throughout the exam, this would produce a very high $\hat{\alpha}$ value; that doesn't mean the test is good, but it is *reliable*.

This idea of testing the correlation of items to the overall exam extends to the item level analysis (*discrimination*). A primitive, yet common, metric is based on splitting the sample based on overall exam performance. For instance, if the overall exam scores are the vector S, we could split S on the median to create two equally sized vectors, S_H and S_L. Specifically looking at item i, $Y_{i|S_H}$ is the percent of students in the high group who correctly answered item i and $Y_{i|S_L}$ is the percent of students in the low group that correctly answered item i. Therefore, the index of discrimination can be stated as:

$$D_i = Y_{i|S_H} - Y_{i|S_L} \tag{5}$$

The split on the median is not the only option. Truman Kelley[4] suggested splitting the sample into three groups: 27% on top and

[2] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. DOI: 10.1007/BF02310555

[3] Cronbach's α appears in the column 'alpha' in Canvas' CSV quiz item analysis.

Sometimes software programs (e.g. ExamSoft) will refer to this score as "KR20." This is because the α estimates are often a special case where the items are dichotomous (no partial credit) – this was first developed by Kuder and Richardson.

G Frederic Kuder and Marion W Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937. DOI: 10.1007/BF02288391

[4] Robert L. Ebel. Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14(2):352–364, 1954. DOI: 10.1177/001316445401400215

bottom and discarding the middle 46%.[5] Of course the problem with this measure is that it isn't really built from a model and it is sensitive to the split point. A more sophisticated metric of discrimination is a version of Pearson's r.[6]

$$r_{pb_i} = \frac{\bar{S}_1 - \bar{S}_0}{\sigma_S}\sqrt{\frac{n_1 n_0}{n^2}} \tag{6}$$

Where $\bar{S}_1$ is the mean score of the students who correctly answered item i, $\bar{S}_0$ is the mean score of students who did not answer item i correctly, and n_1, n_0, and n represents the sample sizes of each group and the sample as a whole. Most often this is referred to as simply r_{pb} or the *point biserial correlation coefficient*.[7] If the question is of high quality, one would expect $r_{pb_i} > 0.2$. An item resulting in a negative r_{pb_i} value indicates that the students who performed the best on the assessment overall actually performed worse on this question. This would indicate that their increased knowledge is leading them to select an incorrect answer. In all likelihood, such a question should be revised.

[5] Canvas follows Kelley's procedure of splitting the sample into three groups.

[6] While many testing software packages include some index of discrimination (D_i), if the platform also includes point biserial correlation coefficient (r_{pb_i}), D_i should probably be ignored as it is sensitive to the split points.

[7] Akindi refers to this as their "discriminatory score." ZipGrade refers to this as "discriminant factor." Quick Key describes this value as "discrimination analysis." Canvas and ExamSoft correctly describe it as the point biserial correlation coefficient.

Item Response Theory (IRT)

CTT's modeling is rudimentary to say the least. While error is understood, how that error is occurring is not really expressed. This results in metrics that are measures of performance rather than latent traits. The IRT story is of increasingly more sophisticated models to correctly estimate latent traits. This results in estimates of *difficulty*, *discrimination*, and *guessing* – many of the same concepts as in CTT. IRT estimation methods are now routinely included in both open source software packages (e.g. R, Python) and commercial packages (e.g. STATA, SAS). IRT estimates are not currently available in testing programs such as Canvas and Akindi. However, especially given the size of some university classes, there is no reason such metrics couldn't be included in the future as the software programs have all the data necessary to calculate these estimates.

The Rasch or One Parameter Logistic (1PL)

IRT models are built on the concept of a logistic regression ($Pr(t) = 1/(1+e^{-t})$); in essence, expressing answering a question correct a matter of probability. Rasch realized that he could model a student correctly answering a particular question as a function of student n's ability (θ_n) and the difficulty of question i (b_i). Therefore, the Rasch model[8] can be stated as:

[8] Benjamin D. Wright. Solving measurement problems with the rasch model. *Journal of Educational Measurement*, 14 (2):97–116, 1977. DOI: 10.1111/j.1745-3984.1977.tb00031.x

$$Pr(X_{in} = 1) = \frac{1}{1 + e^{-(\theta_n - b_i)}} \tag{7}$$

Under this model, ability of the student is explicitly modeled (and estimated) and *difficulty* is estimated from the data. This model, however, assumes that all items are equally discriminating. This, of course, is likely not the case. Therefore, an obvious extension to the 1PL is the Two Parameter Logistic (2PL).

Two Parameter Logistic (2PL)

The 2PL simply adds a discrimination factor (denoted as a_i) to the Rasch model.

$$Pr(X_{in} = 1) = \frac{1}{1 + e^{-a_i(\theta_n - b_i)}} \tag{8}$$

This measure of discrimination is significantly more sophisticated than what we see in CTT. Nonetheless, there is a connection. Lord[9] showed that under the assumption that ability (θ) is normally distributed, a_i is a monotonic transform of the point biserial correlation coefficient (r_{pb_i}). Specifically:

[9] Frederic M Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980

$$a_i \approx \frac{r_{pb_i}}{\sqrt{1 - r_{pb_i}^2}} \tag{9}$$

Three Parameter Logistic (3PL)

The 2PL is the proper estimation strategy when the exam questions cannot be guessed. However, in the context of multiple choice exams, all students have some probability of correctly answering a question. This inflates the ability estimates in the 2PL model for all questions. As the ability/discrimination is specified linearly, both estimates can be incorrect when guessing is a substantial percent of correct responses. Adding a 'guessing parameter' (c_i) solves this problem:

$$Pr(X_{in} = 1) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_n - b_i)}} \tag{10}$$

Given sufficient data, this is the most accurate means to estimate the latent parameters in a multiple choice exam setting; most software programs producing IRT estimates are using this equation. The estimates of a_i and c_i play an important role in diagnosing exam quality. c_i should converge to $1/k$ (e.g. 25%). When it does not, this means the question distractors are either confusing or can be eliminated by some students. a_i can be used to rank the questions from most discriminating to least (a value below 0.2 is suspect, a value above one is excellent).

As there are a number of free parameters in the 3PL, it requires a large number of observations;[10] Han[11] suggests fixing the c_i parameter at $1/k$ to increase the degrees of freedom in smaller datasets. This should encourage instructors to use nationally-normed exams (such as the Test of Understanding in College Economics[12]) as the c_i and a_i are population independent.

[10] R. J. De Ayala. *The theory and practice of item response theory*. Guilford Publications, New York, NY, 2013

[11] Kyung T Han. Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, 17(1):1–24, 2012

[12] William B Walstad, Michael Watts, and Ken Rebeck. *Test of understanding in college economics: Examiner's manual*. Council for Economic Education, New York, 4th edition, 2007

Flow of Knowledge or Value-Added Learning Models

While CTT and IRT testing models are largely concerned with test quality and knowledge, educators are often concerned with learning. The ability factor in the 3PL measures (θ_n) the student's ability/knowledge at the time of the exam. Notably, θ_n does not indicate if the student learned the material in response to a treatment (e.g. class or program) or already knew the material. The simplest, and still the most common, form of value-added learning score is to subtract the pre-test from the post-test (post-test $-$ pre-test). As this technique measures knowledge at two points in time and calculates the difference, it is often called the 'flow of knowledge.'

The pre- and post-test percentages should only include students who took both exams.

An approach adopted by some is to treat the pre-test score as an independent variable in a regression model. For instance, if the vector post-test contains the students' post-test scores and pre-test contains the students' pre-test scores then:

Using a regression model to estimate the impact of teachers, principles, or an entire institution (e.g. the Collegiate Learning Assessment or CLA+) is commonly referenced by the media as value-added learning. However, it is just one specific type and use of value-added learning measurements.

$$\text{post-test} = \alpha + \beta X + \gamma \text{pre-test} \tag{11}$$

Instead of assuming that the pre-test has a unit impact on performance, γ allows the pre-test's impact to vary. This partially controlled for the bounds issue with simply subtracting the pre-test from the post-test: the upper bound is $1 -$ pre-test. Hake[13] addressed this issue more directly by linearly transforming the difference metric:

$$\frac{\text{post-test} - \text{pre-test}}{1 - \text{pre-test}} \tag{12}$$

[13] Richard R Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74, 1998. DOI: 10.1119/1.18809

The Hake gain score is attractive as it re-scales the learning value to a $[0, 1]$ range; it has been widely adopted in the hard sciences. However, it treats all learning types as the same. A more recent approach suggested by Walstad and Wagner[14] disaggregates the value-added learning values into four types: positive learning ($\hat{\text{pl}}$), negative learning ($\hat{\text{nl}}$), retained learning ($\hat{\text{rl}}$), and zero learning ($1 - \hat{\text{pl}} - \hat{\text{nl}} - \hat{\text{rl}}$). Positive learning occurs when a student who did not know the material learned the material in the course of the class. Negative learning occurs when a student forgets the material in the course of the class. Retained learning indicates the student knew the material both times. Zero learning indicates the student never knew

[14] William B Walstad and Jamie Wagner. The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education*, 47(2):121–131, 2016. DOI: 10.1080/00220485.2016.1146104

the material. Figure 1 describes the mapping between student pre- and post-test performance and the four learning types.

	Correct (Post)	Incorrect (Post)
Correct (Pre)	$\hat{\text{rl}}$	$\hat{\text{nl}}$
Incorrect (Pre)	$\hat{\text{pl}}$	$1 - \hat{\text{pl}} - \hat{\text{nl}} - \hat{\text{rl}}$

Figure 1: Value-added disaggregation as described in Walstad and Wagner (2016).

The method outlined by Walstad and Wagner works well in the context of exams where guessing has a near-zero probability of success. However, in the context of multiple choice exams, *learning performance* is a function of both a latent trait and guessing.[15] Specifically, γ is true positive learning (adjusted for guessing), α is true negative learning (adjusted for guessing), and μ is true stock knowledge at the time of the pre-test (adjusted for guessing); true retained learning is $\mu - \alpha$. Smith and Wagner developed estimates of these latent traits when properly accounting for guessing.

[15] Ben O Smith and Jamie Wagner. Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores. *The Journal of Economic Education*, 49(4):307–323, 2018. DOI: 10.1080/00220485.2018.1500959

$$\hat{\gamma} = \frac{\hat{c}_i(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{(\hat{c}_i - 1)^2} \tag{13}$$

$$\hat{\alpha} = \frac{\hat{c}_i(\hat{\text{pl}} + \hat{\text{rl}} - 1) + \hat{\text{nl}}}{(\hat{c}_i - 1)^2} \tag{14}$$

$$\hat{\mu} = \frac{\hat{\text{nl}} + \hat{\text{rl}} - \hat{c}_i}{1 - \hat{c}_i} \tag{15}$$

Where the raw learning values are as defined by Walstad and Wagner and $\hat{c}_i$ is the probability of guessing correctly. In the context of nationally-normed exams, $\hat{c}_i$ can be estimated using the 3PL and substituted into the above equations. Otherwise, it is usually assumed that $\hat{c}_i = 1/k$ where k is the number of question options.

In terms of instruction/exam diagnostics, $\hat{\alpha}$ plays an important role. Ideally, $\hat{\alpha}$ should converge on zero (similar to how c_i should converge on $1/k$ in the 3PL model). When it does not, this could be occurring for one of three reasons:

There is a convergence between Smith and Wagner (2018) and Hake (1998). Smith and White (2021) show that $\hat{\gamma}/(1 - \hat{\mu})$ converges to the Hake gain estimator when $\hat{\alpha} = 0$. The Hake estimator is, therefore, a special case of the family of guessing-adjusted estimators.

1. There is an insufficient number of observations and the practitioner is seeing statistical noise.
2. $\hat{c}_i$ was assumed to equal $1/k$ where, in truth, it does not. This indicates that the exam question is likely suspect and should be redesigned.
3. The instruction of the underlying content was confusing enough to result in some students 'un-learning' the material. These lesson plans should be redesigned.

Smith and White[16] introduces a gain transformation ($\hat{\gamma}/(1 - \hat{\mu})$) of the above guessing-adjusted estimators and sensitivity test to determine if the transformed form is more robust.

[16] Ben O Smith and Dustin R White. On guessing: An alternative adjusted positive learning estimator and comparing probability misspecification with monte carlo simulations. *Applied Psychological Measurement*, 45(6):441–458, 2021. DOI: 10.1177/01466216211013905

$$R = \frac{\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1}{2\hat{\text{pl}} + (\hat{\text{nl}} + \hat{\text{rl}} - 1)(\hat{c}_i + 1)} \tag{16}$$

The gain transformation is less sensitive to probability misspecification when R is between $[-1, 1]$. Outside of that range, the original $\hat{\gamma}$ estimator is more robust.

The Assessment Disaggregation software [Smith, 2022] available at https://www.assessmentdisaggregation.org/ estimates the learning values developed by Walstad and Wagner (2016), Smith and Wagner (2018), and Smith and White (2021) using standard exam/quiz files.

Rubric-Based Knowledge Measures

The above section discussed techniques to measure learning using a pre- and post-test. However, many times students are assessed using projects, presentations, or theses. These are typically graded using a set of rubric rows. Smith and Wooten[17] showed that parameters that characterize the rubric-row difficulty and student's ability can be estimated. Specifically, they transform the problem into a special form of survival function. Consider the following rubric row:

[17] Ben O Smith and Jadrian J Wooten. Assessing proxies of knowledge and difficulty with rubric-based instruments. *Southern Economic Journal*, 90(2): 510–534, 2023. DOI: 10.1002/soej.12658

Did not meet any of the requirements	The chosen topic was on-topic for the course, but the literature was lacking	The literature review was complete but contained errors	Full credit

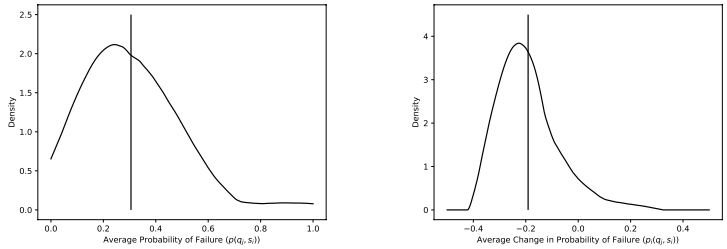
For the student to achieve a score in the third box from the left, they must have met the requirements of the second box from the left. Similarly, a student achieving the box on the far right indicates they met all of the requirements for all other boxes.[18] Therefore, one can express the problem as a probability to fail to achieve the next box based on rubric-row difficulty and student ability. This probability is expressed as $p(q_j, s_i) = 1/(1 + e^{-(q_j + s_i)})$, where q_j is rubric row difficulty and s_i is student ability. While it is difficult to interpret the estimates for q_j and s_i directly, one can interpret these estimates by converting them to the Average Probability of Failure (also known as Average Logistic) or Average Change in the Probability of Failure (also known as Average Marginal Logistic). Consider the rubric row results below:

[18] The box on the far-right is a special case as the data can be top censored. The estimation procedure proposed in the paper takes this into account, but it is outside the scope of the of this handout.

Rubric Variable	E. Value	Average Logistic	Average Marginal Logistic
1	-1.736	0.073	-0.100
2	0.713	0.383	0.139
3	1.215	0.483	0.246

For instance, examining rubric row 1 reveals an estimated value of q_j of -1.736. Looking at the columns Average Logistic and Average Marginal Logistic we can see how this estimate translates into probability. On average, students had a failure rate of about 7% on rubric row 1 and, all else equal, rubric row 1 was about 10% easier. By contrast, rubric row 3 was hard. On average, students had a failure rate of about 48% and it was about 25% harder, all else equal.

Student ability s_i can be similarly examined but it is probably more meaningful to examine the entire distribution. For instance in a large Labor Economics class, the Average Probability of Failure and Average Change in the Probability of Failure can be seen in the below figure.



(a) Probability of Failure (b) Change in Probability of Failure

These distributions of student ability can be used to measure ability of one group of students in comparison to others (e.g. time trend data, treatment effects) or comparing the same students' ability at two different times.

The Project Based Assessment web application at https://projectassessment.app implements the estimation technique proposed by Smith and Wooten.

Conclusion

Exam software routinely claim to provide exam analytics. Unfortunately, the analytics provided have often been superseded by more advanced methods. Fortunately, the more advanced methods, such as IRT, are built into most statistical packages and Value-Added Learning Models and Rubric-Based Knowledge Measures are available in open source statistical packages (R, Python).

References

Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. DOI: 10.1007/BF02310555.

R. J. De Ayala. *The theory and practice of item response theory*. Guilford Publications, New York, NY, 2013.

Robert L. Ebel. Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14(2):352–364, 1954. DOI: 10.1177/001316445401400215.

Richard R Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74, 1998. DOI: 10.1119/1.18809.

Kyung T Han. Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, 17(1):1–24, 2012.

G Frederic Kuder and Marion W Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937. DOI: 10.1007/BF02288391.

Frederic M Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980.

Ben O Smith. Assessment disaggregation: A new tool to calculate learning types from nearly any exam platform, including online systems. *The Journal of Economic Education*, 53(2):194–195, 2022. DOI: 10.1080/00220485.2022.2038321.

Ben O Smith and Jamie Wagner. Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores. *The Journal of Economic Education*, 49(4):307–323, 2018. DOI: 10.1080/00220485.2018.1500959.

Ben O Smith and Dustin R White. On guessing: An alternative adjusted positive learning estimator and comparing probability misspecification with monte carlo simulations. *Applied Psychological Measurement*, 45(6):441–458, 2021. DOI: 10.1177/01466216211013905.

Ben O Smith and Jadrian J Wooten. Assessing proxies of knowledge and difficulty with rubric-based instruments. *Southern Economic Journal*, 90(2):510–534, 2023. DOI: 10.1002/soej.12658.

Ross E. Traub. Classical test theory in historical perspective. *Educational Measurement: Issues and Practices*, 16(4):8–14, 1997. DOI: 10.1111/j.1745-3992.1997.tb00603.x.

William B Walstad and Jamie Wagner. The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education*, 47(2):121–131, 2016. DOI: 10.1080/00220485.2016.1146104.

William B Walstad, Michael Watts, and Ken Rebeck. *Test of understanding in college economics: Examiner's manual*. Council for Economic Education, New York, 4th edition, 2007.

Benjamin D. Wright. Solving measurement problems with the rasch model. *Journal of Educational Measurement*, 14(2):97–116, 1977. DOI: 10.1111/j.1745-3984.1977.tb00031.x.