99.8% Reliable Assessment with CyberTutor

Elsa Sofia Morote & David E. Pritchard
Massachusetts Institute of Technology
Cambridge, MA 02139.

1. Introduction

Traditionally teaching and testing (i.e. assessment of what students know) are separate activities. CyberTutor, designed as an interactive tutor, can integrate tutoring and assessment, which will improve both activities. The following scenario illustrates the basic idea:

Imagine that a rich ship-owner has hired Socrates to tutor his children. At the end of the month he desires to assess the amount they have learned. Would you advise him to:

a. Administer a standardized hour-long test to the children?
b. Ask Socrates how much they have learned?

This multiple-choice question is rhetorical. Clearly Socrates' hours of individual interaction enables him to form a far more accurate and nuanced assessment of each student than could be provided by a one-hour test. Moreover, this *integrated assessment* is *continuously updated* and is used by Socrates to adjust the pace and content of the session. Importantly, this continuous integrated assessment is available without the stress and time lost of traditional testing.

Like Socrates, CyberTutor can integrate instruction and assessment. The following data shows that this assessment has about 62 times less variance due to random test error than a three-hour final examination in a typical physics course. This huge improvement results from the combination of two factors. First, the typical student interacts with CyberTutor for at least 48 hours per term, 16 times longer than the test, diminishing the

effect of lucky guesses and careless errors. Secondly, by considering requests for hints, solutions, and importantly the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a much better determination of the student's skill on a problem. This is particularly true of a problem that is too difficult for the student as presented – CyberTutor's hints and spontaneous responses to wrong answers systematically adapt the difficulty of the problem down to the student's skill level, and the CyberTutor assessment algorithm accurately determines that level.

The CyberTutor analog of Socrates' assessment is a **skill profile** – that student's skill rating on each of a set of predetermined topics. A topic could be an item in the regular syllabus like momentum, but it could also be a foundational skill like vector components, a general skill like using dimensions and units, or a conceptual topic like Newton's concept of force.

Because it is continuously updated, skill profiling has the following potentially revolutionary benefits for the student and the teacher

1) It is vastly fairer to use as the basis of a grade because of its dramatically smaller random error
2) Students who have gaps in their foundational skills can be identified and helped early in the course
3) The skill level of the class on the current topics can guide the teacher's allotment of class time – a new form of Just In Time Teaching.
4) The assessment is sufficiently detailed to be used to predict the student's grade on a standard test, for example a midterm or the AP physics test.
5) Teachers can therefore dispense with some examinations and increase class time

Skill profiling can have a revolutionary impact if it is used as a grade (e.g. your grade on the momentum homework is your skill at momentum on Thursday night at 10PM). This could replace the students' current strategy of "avoid lost points" with the much improved one "learn the most."

2. Reliability of Assessments

By definition, someone taking a test or other assessment measure achieves the "observed score." The observed score is thought to be composed of two different components: The true score – an exact measure of the amount of the trait or skill assessed, offset by a random error score. The error may result from individual student responses to particular questions (Manson & Branle, 1997) or external conditions such as fatigue and noise. Thus,

Observed score = True score + Error score

The variance of the observed score is composed of the variance of true scores and the variance of error. Then,

$$\sigma^2_o = \sigma^2_t + \sigma^2_e$$

where, σ^2_o = variance in observed scores, σ^2_t = variance in true scores, and σ^2_e= variance of error

The "reliability" of a test or assessment measures its repeatability or freedom from random testing errors. An assessment is considered reliable if it would give us the same result over and over again. By definition, the reliability is the ratio of variance of the true skill divided by the observed variance of the measurement (observed variance).

$$\text{Reliability} = [\sigma^2_t] / [\sigma^2_o] = [\sigma^2_t] / [\sigma^2_t + \sigma^2_e], \text{ so}$$
$$\text{Reliability} = 1 - \sigma^2_e / \sigma^2_o$$

The reliability provides a direct guide to the reproducibility of the grade between two similarly prepared tests (odd or even numbered questions) on equivalent material. The second line shows that as the reliability => 1, the error => 0.

A standard method of determining the reliability of an assessment is to divide the test questions into two subtests, for example the odd and even questions if they were equal in coverage of topics and difficulty. If the resulting subtests determine skill with perfect reliability, each student would receive the same score on both subtests, and a scatter plot of scores on the "odd (a)" and "even (b)" subtests would lie on the diagonal. We compute the reliability of the various assessments considered here from the uneven split formula derived in the appendix. This method reduces to the usual split half reliability formula when both (a) and (b) halves are equally difficult, and does not assume that the halves have equal standard deviations, an additional assumption of the Spearman-Brown reliability formula[1] (See Appendix for derivations).

$$\text{Reliability} = 1 - (1+\beta)^2 / 4\beta * \sigma^2_{(a-b)} / \sigma^2_o + (1-\beta)^2 / 4\beta = 0.85$$

where $\beta = \mu_a / \mu_b$, and $\sigma^2_o = \sigma^2_{(a+b).}$ In this formula, the key determiner of the reliability is the mean variation of the difference between the scores on the even and odd collections of problems. Even if the mean on the "a" questions is 20% higher than that on the "b" questions, the error σ^2_e equals $\sigma^2_{(a-b)}$ to within 1%, and the last term changes the reliability by less than 1% also.

2.1. Three Hour Final Examination

The split test reliability method was applied to the MIT 8.01 final exam for 2001 – see Fig. 1. To the eye, there is considerable lack of consistency between the performances on the two subtests. This reflects the error on the two subtests – they covered the same material, but random errors are considerable. Of course, adding the two halves together reduces the relative size of the error and results in a reliability = 0.85 as computed from the unequal split halve formula above.

[1] Spearman-Brown formula: Reliability = 2 * [correlation (Odd, Even)] / [1 + correlation(Odd, Even)]

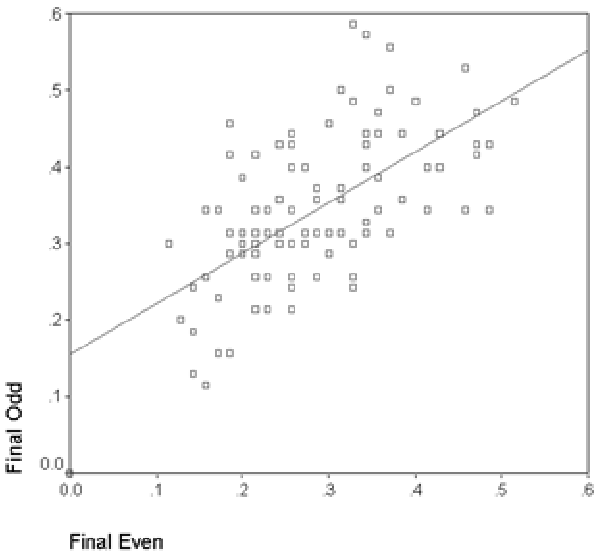

Figure 1. Scatter plot of the grades of each individual student on the two subtests comprising the MIT 8.01 Final Exam in 2001. For these points, $r^2 = 0.41$.

The reliability of this test – 0.85 - is a moderate reliability in a test. Generally reliability above 0.90, typical of ETS tests, is considered acceptable for a test in education and behavioral sciences (Manson & Bramle, 1997). This reliability means that 85 percent of the variance in the observed scores is due to true score variance. The error causes 15 percent of the observed variance. Taking the square root yeilds $\sigma_{error}/\sigma_{observed} = 0.38$, showing that the error is a substantial fraction of the width of the observed distribution. Such large error can lead to a misdetermination of who should pass and fail (see "Fairness" in section 4.1).

We note in passing that this particular final exam had three features that should give it less error than typical physics finals. We used a standard test (Mechanics Baseline Test) for 1/3 of the test, we divided the remainder into basic skill problems and complex problems which accentuate the performance of different students, and the total test requested 47 specific responses from the students – in comparison, the 8.01 final in the preceding semester required only 28, and therefore probably had larger relative statistical fluctuations.

2.2. CyberTutor Lost Points Assessment

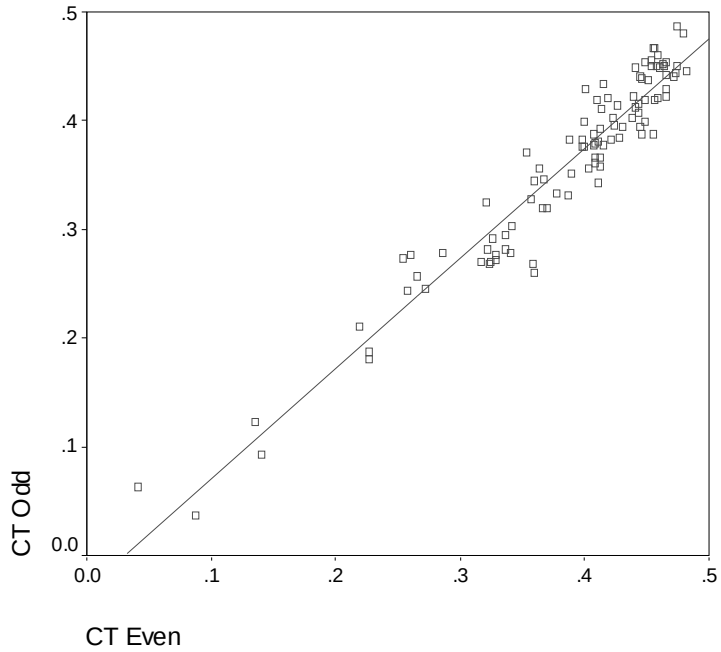

Figure 2. Scatter plot of Lost Points Algorithm applied to Even versus Odd CyberTutor problems in spring 2001. r^2 =0.93.

In spring 2001, CyberTutor was graded using an ad hoc *lost points algorithm*, whereby the grade was 1 for each main part of any problem with a small penalty for requesting hints and a significant penalty for requesting any solution, even for the subparts. The reliability = 0.977 was found from the uneven split test formula.

Although the algorithm is rather crude, this assessment has considerably less error than the final examination as shown in Fig. 2. The improvement results from the fact that students spent about 16 times as much time doing CyberTutor over the term as taking the final. Since the error variance is reduced by only 6.48 times, we are forced to conclude that the lost points algorithm is only 6.48/16 = 41% as error-free as the final exam *per unit time.* This probably reflects the slower rate of work of students on their homework as opposed to their final examination, together with the ad hoc nature of the algorithm.

2.3. Enhanced CyberTutor Algorithm

If someone is highly skillful, they generally make fewer errors and take less time to successfully complete a task than a less skillful person. Moreover, any individual has the option to work more quickly at the expense of making more errors. Thus if we include the number of wrong responses and the time taken by students doing CyberTutor problems, we get much more insight into their intrinsic skill level than with the previously discussed "lost points" algorithm. With these considerations in mind, we developed an assessment algorithm which mimics the lost points grade, but depends on these additional variables (time and wrong answers) as well as all other variables available from the log of each student's interactions with CyberTutor including among others, hints requested, solutions requested, number of problems worked, etc. We call this assessment algorithm the *Enhanced CyberTutor Algorithm*

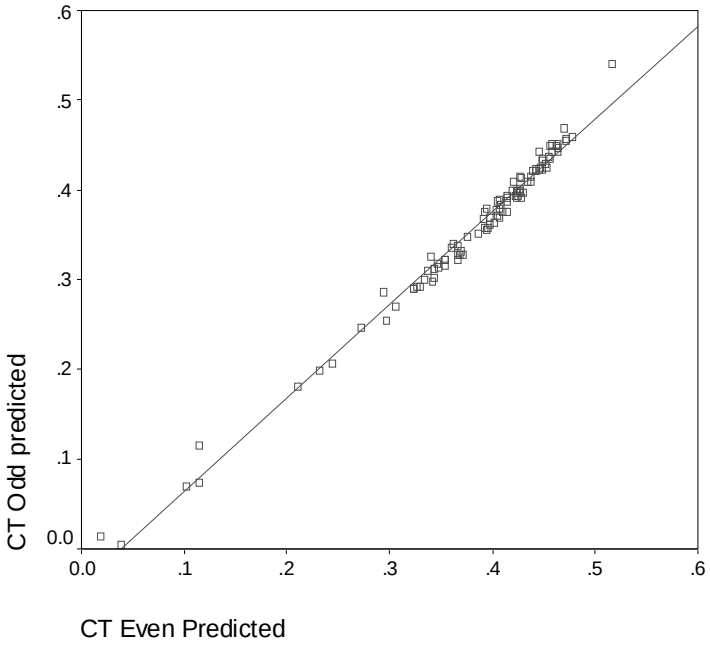

Fig. 3. Scatterplot of the Enhanced CyberTutor Algorithm as computed from the even problems and the odd problems for each student. The curvature indicates a small excess of easier problems in the even group.

The Uneven Split reliability of the combined assessment is 0.9976. This is a fantastic improvement over the final examination. That means that 99.8 percent of the observed variance in scores is due to true score variance. The CT enhanced algorithm has 4 times less error *per unit time* than the final exam. Overall, it has 62 times less error than the final exam. It provides nearly error-free assessment which as far as we know is the world record for high reliability for any assessment in education. The implications of this are discussed in the last section of this paper.

2.4. Summary

Table 1 shows the differences among Final Exam, CT Lost points, and Enhanced CT. The Final exam has lower reliability than the C-T lost points algorithm, and far less than the enhanced CT assessment. The key finding is that enhanced CyberTutor assessment gives 62 times less error than the 3-hour timed examination.

Table 1. Reliabilities' Summary

	r^2	Reliability	σ^2_e / σ^2_t	σ_e / σ_t
Final Exam	0.41	0.85	0.17	0.42
CT Loss points	0.93	0.977	0.023	0.15
Enhanced CT	0.986	0.9976	0.0024	0.05

This huge improvement in the variance of the error results from the combination of two factors. First, the typical student interacts with CyberTutor for at least 48 hours per term, 16 times longer than the test, diminishing the effect of lucky guesses and random errors. Secondly, by considering requests for hints, solutions, the number of wrong answers given, and importantly the time the student takes to complete each part of the problem, CyberTutor can make a much better determination of the student's skill per unit of interaction time.

3. Predictive Validity

Validity relates to the meaning of the scores and the ways we use scores to make decisions (American Psychological Association, 1985). Validity addresses the issue of how well do these scores reflect the physics achievement. One way to empirically evaluate validity of the CyberTutor data is so study its "predictive validity", its ability to predict some subsequent performance. The correlation coefficient between prediction and subsequent performance now becomes the validity coefficient. The skill profile can be used to predict each student's grade on a standard test. As a first step in this direction, we developed an algorithm to predict students' grades on the "adjusted" MIT final exam. The adjustment reflected the findings from item analysis that both exam questions concerning angular dynamics were poor measures of overall skill. It is well known that many students "hit the wall" of misunderstanding here, and while we believe that an examination of CyberTutor data concerning angular dynamics would improve our predictive ability on these problems, it was cleaner to eliminate them for this study of validity.

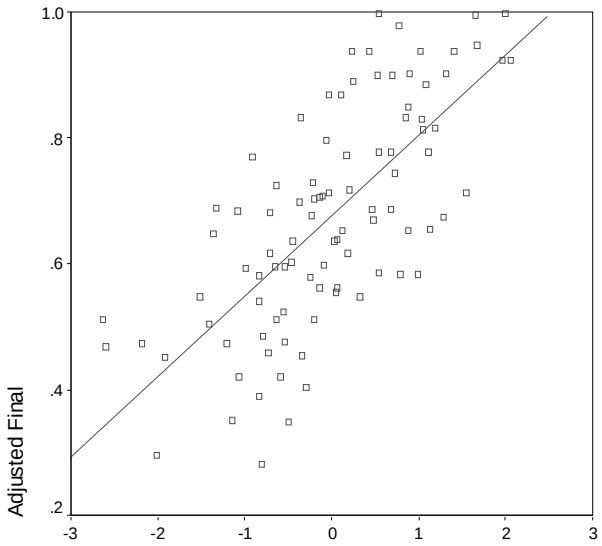

Figure 4. Prediction of scores on Adjusted Final Exam from CyberTutor data, r^2=0.51. The horizontal axis is deviation from CT prediction in standard deviations

The final prediction (Figure 4) yields a value for r, the predictive validity coefficient, of 0.71, which is highly statistically significant[2]. In fact this correlation coefficient would be higher were in not for the 15% unreliability of the final exam. This suggests that the "real" validity coefficient is around 0.77. This level suggests that a variant of the "Enhanced CyberTutor algorithm" would enable us to predict the students' skills with less error than a one-hour test such as a midterm could determine it.

We are optimistic that a prediction based on a detailed skill profile instead of the overall performance could do considerably better at predicting the final exam scores and conclude that CyberTutor as a measurement instrument is not only extremely reliable but also has significant validity. Indeed there is every reason to believe that it would give a better indication of students in trouble at midterm time than the midterm itself.

In this section, we have begged the question of the validity of adopting the final exam as a criterion of validity (e.g. only about 1/4 of the final exam involved conceptual questions). In the future we hope to use the performance in the subsequent required physics course as another measure of validity.

[2] The minimum value of r required to be statistically significant at the 0.05 level a sample size between 100 and 1000 is 0.197 (Hopkins and Glass, 1978 p. 409).

4. Implications of Enhanced Assessment

4.1. Fairness

The final exam contains so much error that 1/4 of the students assigned failing grades have a passing skill, and an equal number of students who should have failed are assigned a passing grade. Thus, for each two students failed, one student was misdiagnosed.

To see how this arises, look at Figure 5a where failure is set at one standard deviation below average, which fails 16 percent of the students, the typical failure rate for 8.01. The dashed line shows the true skill distribution inferred from the measured error and the observed skill distribution (heavy line), which is slightly broader. The number of students failed is indicated by the small curve centered on skill = 40. (The width of this curve is fortuitously very nearly the width of the error, that is 38% of σ_{true}.) Below skill = 35 the curve of students failed lies on top of the true skill distribution, indicating that all students whose skill is this low fail. Between 30 and 46 (the selected true skill for passing) it drops well below the true skill distribution, indicating that some students (lighter shaded area) whose ability is below passing will pass. The distribution of failed students does not drop immediately to zero above true skill = 46, indicating that some students (darker shaded area) whose true ability is passing will fail the test.

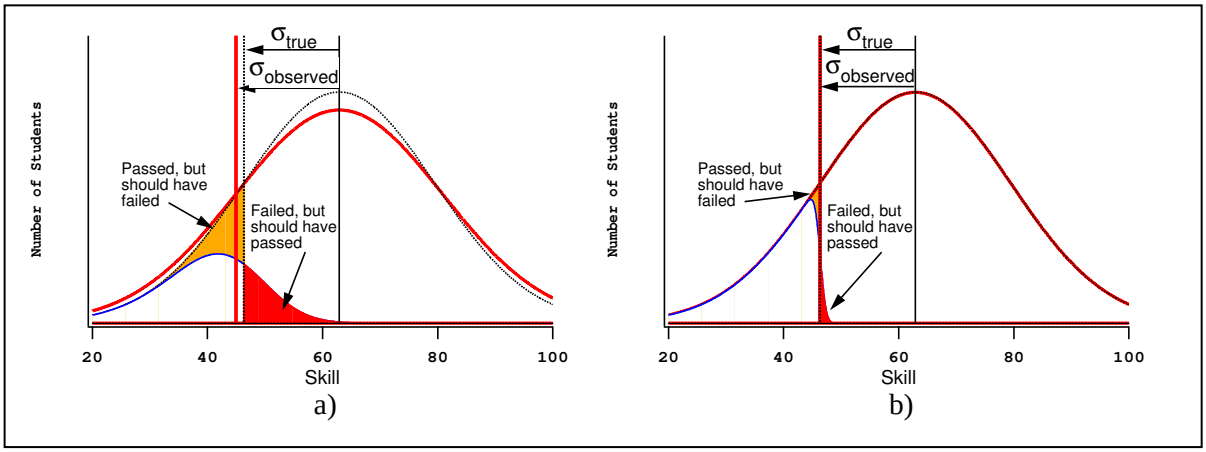

Figure 5. Comparison of CyberTutor and Final Exam Fairness
In both plots, the the dashed curve is the distribution of true skill and the thick curve is the distribution of observed skill (i.e. grade). The shaded area shows students who fail/pass although their true skill level is passing/failing. **a**) In the case of the final exam, 1 in 4 students who fail actually deserve to pass! **b**) If the much more accurate Enhanced CyberTutor Assessment were used instead of the final exam; only 3 percent of the students failed would be misdiagnosed. In a class of 100, only one student would be misdiagnosed, and his true skill would almost certainly lie only 1 or 2 points of the pass-fail line.

4.2. Detailed Assessment and Remediation

The CyberAssessment has such low error that it is capable of providing accurate assessments on a large number of pre-selected topics – thus allowing the detailed skill profile discussed in the introduction. If we took 20 topics – say the major topic of the 13 weeks of the course plus 7 foundational topics; the assessment on each topic would be three times as reliable as the overall assessment of the current final. Even if we increased the skill resolution to include each of the major weekly topics (i.e. perhaps 180 topics) the variance on each would be small enough to identify those students requiring immediate assistance on a particular topic. By basing the assessment on a more careful analysis of wrong answers (was some particular step omitted or a particular wrong assumption made, or was this just a careless error?) we project that we can improve the CyberTutor assessment sufficiently to measure even this fine grid with the accuracy that the current final determines the overall skill of the students.

Such a fine-grained assessment would permit the identification of skill deficiencies in foundational skills as each student moved through the course. This would allow tutors (or perhaps additional CyberTutor problems) to remedy the deficiency before it causes irreparable harm to the student. Additionally, knowing the class' skill profile on each detailed topic could allow the teacher's next lecture to review topics that were not being well learned.

4.3 Powerful Education Research

The unprecedented reduction of assessment error demonstrated above provides educational researchers with an assessment of unprecedented precision. CyberTutor error is currently 0.05 σ^2_o, or about 0.02 of the average score on a per student basis, diminished to 0.006 for a class with 100 students. Since we have seen 50% changes in score on one problem due to the preceding one, we can measure the educational effect of a single educational item (such as reading a subchapter of a book or viewing a demonstration) on the students' subsequent performance with exquisite precision – certainly enough to make highly significant comparisons of two comparable educational items. All sorts of educational items can be tested and improved this way.

References

American Psychological Association (1985). Standards for educational and psychological testing. Washintong, DC: The Association

Kulik, C., Kulik, J., & Bangert-Drowns R. (1990). Effectiveness of Mastering learning programs: A Meta Analysis. Review of Educational Research 60, 265-299

Hopkins, K., Stanley, J., & Hopkins, B.R. (1990). Educational and Psychological Measurement and Evaluation.

Hopkins, K., & Glass, G. (1978). Basic Statistics for the behavioral sciences (end Ed). Englewood Cliff, NJ: Prentice-Hall

Lesgold, A., Lajorie, S.P., Bunzo, M.& Eggan, G. (1992). A coached practice environment for an electronics troubleshooting job. In Larkin, R. Chabey, & C. Chetic (Eds.) Computer assisted instruction and intelligent tutoring systems: Establishing communication and collaboration (pp. 201-238). Hillsdale, N.J. Lawrence Elbaurm Associates.

Shute, V., & Glazer, R. (1990). Large scale evaluation of an intelligent discovery world: Smithtown. Interactive Learning environments, 1 51-77.

Manson, E., & Bramle, W. (1997). Research in Education. Sydney:Brow & Benchmark Publishers.

Pritchard, D., & Morote, E. (2001). What Course Elements Correlate with Improvement on Tests in Introductory Newtonian Mechanics? NARST Annual Meeting
New Orleans, Louisiana, April 7-10, 2002.

Regian, J., & Shute, V. (1990). Cognitive Approaches to automated instruction. Lawrence Erlbaum Associates, Publishers.

Slavin, E. (1987). Mastering Learning Reconsidered. Review of Educational Research, 57, 175-213

Wackerly, D. , Mendenhall, W., Scheaffer, R. (1999). Mathematical Statistics with Applications. Belmot: Duxbury Press

Appendix

Uneven Split Reliability test – case of uneven splits

Computing reliability by using the approach of equivalence, where two equivalent forms of the test are formulated and administered to the same persons has a major problem, which is in developing two tests that are perfectly equivalent. It can also be applied splitting a test into halves. The Uneven Split reliability test is proposed to overcome this problem.

Proposition: If we split an exam in two A and B uneven halves, that is $\mu_a = \beta\mu_b$, where μ is average, and consider that the errors $\sigma^2 a_{error}$, $\sigma^2 b_{error}$ of each half are not the same, then the Uneven Split reliability test will be

$$\text{Uneven Split Reliability} = 1 - (1+\beta)^2/4\beta * \sigma^2_e/\sigma^2_o + (1-\beta)^2/4\beta$$

Proof. Suppose a Final Exam is divided into two equivalent halves, A and B, then the average score on the final will be equal to the sum of the average of part A and part B

$$\mu_{final} = \mu_a + \mu_b$$

If the halves are perfectly equivalents then $\mu_a = \mu_b$. However, in most of the cases the halves cannot be split evenly, then

$$\mu_a = \beta\mu_b \text{ and } \sigma_a = \beta\sigma_b, \text{ then } \beta = \mu_a / \mu_b \qquad (*)$$

The observed score is the score that someone gets on a test or measurement. The observed score may be though as of results of two different components: The amount of the trait possessed, called the true score, which is not influenced by random or external conditions such as fatigue or noise from outside the room. The other component reflects the effects of these external chance conditions and is called the error score (Manson & Branle, 1997). Thus, if $X_{observed}$ represents the observed score, X_{true}, the true score and X_e the error score, then

$$X_{observed} = X_{true} + X_{error}$$

and if σ^2_o represents the variance of observed score, σ^2_t , the variance of true score and the variance of error, σ^2_e then,

$$\sigma^2_o = \sigma^2_t + \sigma^2_e \qquad (1)$$

The reliability coefficient is defined as the ratio of the variance of true scores to the observed scores (Mason & Branle, 1997), then

$$\text{Reliability} = \sigma^2_{true} / \sigma^2_{observed} = [\sigma^2_{true}]/ [\sigma^2_{true}+\sigma^2_{error}]$$
$$= 1- \sigma^2_{error} / \sigma^2_{observed} \quad (2)$$

From (1), the variance of each half will be

$$\sigma^2 a_o = \sigma^2 a_t + \sigma^2 a_e \quad (3)$$

$$\sigma^2 b_o = \sigma^2 b_t + \sigma^2 b_e \quad (4)$$

It follows that σ^2_o of the test will be $\sigma^2_{(a+b)o}$, where

$$\sigma^2_{(a+b)o} = \sigma^2 b_t + \sigma^2 a_t + 2\, \sigma (a, b)_t + \sigma^2 a_e + \sigma^2 b_e$$

From (*) we know that $\sigma_a = \beta\sigma_b$ then

$$\sigma^2_{(a+b)o} = \sigma^2 b_t + \beta^2\sigma^2 b_t + 2\, \beta\, \sigma b_t + \sigma^2 a_e + \sigma^2 b_e$$

$$\sigma^2_{(a+b)o} = (1+ \beta)^2\sigma^2 b_t + \sigma^2 a_e + \sigma^2 b_e \quad (5)$$

Similarly,

$$\sigma^2_{(a-b)o} = (1- \beta)^2\sigma^2 b_t + \sigma^2 a_e + \sigma^2 b_e \quad (6)$$

Multiplying equations (5) and (6) by $(1-\beta)^2$ and $(1+\beta)^2$ respectively and subtract them, we have

$$\sigma^2 a_{error} + \sigma^2 b_{error} = [- (1- \beta)^2\sigma^2_{(a+b)o} + (1+ \beta)^2\sigma^2_{(a-b)o}] / 4\beta \quad (7)$$

Applying (7) in (2) we obtained the Uneven Split reliability test

$$\text{Uneven Split Reliability test} = 1- (1+ \beta)^2\sigma^2_{(a-b)o} / (4\beta\, \sigma^2_o) + (1- \beta)^2/ 4\beta$$

$$= 1- (1+ \beta)^2 / 4\beta * \sigma^2_e / \sigma^2_o + (1- \beta)^2/ 4\beta \quad (8)$$

If $\beta = 1$, the formula (8) is reduced to the Reliability test formula on (2):

$$\text{Reliability Test} = 1- \sigma^2_e / \sigma^2_o$$