Lengthening a Test Removes Only Chance Error, Not the Gap in What Is Measured

— raising a test of reliability **0.5** to **0.99** requires **99.000000** times its length, and what grows in proportion to length is not the reliability but its odds — however long the test, the observed correlation with what is wanted tops out at the true **0.6** — [Paper 580]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— From thinking to syncing. —

Abstract

When a test's reliability is low, the advice is "add more questions".What this paper shows is that adding questions removes only chance error; the gap between what is measured and what is wanted is not reduced at all by length, and the correlation with what is wanted tops out at the true correlation.No new theorem or law is claimed. First, the length needed for higher reliability rises steeply——from 0.5 to 0.9 takes 9.000000 times the length, and on to 0.99 takes 99.000000 times in all (Section 2). **Second, reliability never reaches 1——10000 times the length gives 0.99990001, short of 1 by 0.00009999** (Section 3). **Third, what is exactly proportional to length is the odds——the odds of reliability $\rho/(1-\rho)$ are 2.000000 at twice the length, 9.000000 at nine times and 100.000000 at a hundred times, equal to the factor of length** (Section 4). **Fourth, and this is the core.The correlation with what is wanted tops out at the true correlation——with a true correlation of 0.6, the observed correlation is 0.424264 at the original length, 0.569210 at nine times and 0.597022 at a hundred times, and never exceeds 0.6** (Section 5). **Fifth, if the measure of what is wanted has error too, the ceiling drops further——with its reliability at 0.7, the ceiling is 0.501996** (Section 5). **Sixth, the separator is chance error against a gap in what is measured——length removes only the first** (Section 6). **Keywords**: classical test theory, reliability, Spearman-Brown formula, attenuation of correlation, validity.

Scope of this paper (scope note): No new theorem or law is claimed——the Spearman--Brown formula and the formula for attenuation of correlation are standard in classical test theory.Parallel questions are assumed——the added questions measure the same thing with the same precision as the originals; adding questions of a different kind changes what is measured.A true correlation of 0.6 and a reliability of 0.7 for the other measure are representative values——which exam has which values is not stated.Fatigue and practice of the test taker are not included——actually making a test 99 times longer would itself change what it measures.No exam is designed——how long a test should be is not stated.Relation to earlier papers: Paper 334 showed that combining independent measurements adds not values but precisions——"the odds are exactly proportional to length" here is a restatement of that and shares its root. What this paper adds is what lies beyond: combining removes only chance error.Paper 470 touched on reliability in its scope note and showed that a test ninety-nine per cent accurate can be wrong nine times in ten——the shared shape is that stating a test's quality as one number drops what it is good for.Paper 300 showed that sharing a root is decidable——"lengthening raises reliability" and "correlation tops out at the true correlation" have distinct roots.What is added is **setting out the lengths needed for target reliabilities and confirming that they return the targets exactly**, **showing that reliability never reaches 1 while its odds are exactly proportional to length**, **finding that the correlation with what is wanted tops out at the true correlation of 0.6, and drops to 0.501996 when the other measure has error**, and **placing the separator on the distinction between chance error and a gap in what is measured**.

1. Notation and the Question

Setting: let a test score X be the sum of a true score T and a chance error E.The reliability is $\rho = \mathrm{Var}(T)/\mathrm{Var}(X)$.Lengthening the test n times with questions that measure the same thing with the same precision gives

$$\rho_n = \frac{n\rho}{1+(n-1)\rho}, \qquad \frac{\rho_n}{1-\rho_n} = n\,\frac{\rho}{1-\rho}.$$

The observed correlation with what is wanted, Y, is thinned from the true correlation r_{TY} to

$$r_{XY} = r_{TY}\sqrt{\rho_X\,\rho_Y}.$$

> **Observation 1.1**
>
> when a test's reliability is low, the advice is "add more questions".That is right——the question is what adding questions removes and what it does not.

Question: when a test is lengthened, what is reduced and what is not?

2. Length and Reliability

Computation 2.1: **the factor of length needed for a target reliability.**

reliability	length needed
0.5 to 0.9	**9.000000** times
0.7 to 0.9	**3.857143** times
0.8 to 0.9	**2.250000** times
0.5 to 0.99	**99.000000** times

> **Observation 2.2**
>
> **at those lengths the reliability returns exactly to the target (all four to 10^{-12}).From 0.5 to 0.9 takes nine times, but from there to 0.99 takes eleven times more——the last digits are the most expensive.**

3. Never Reaching One

Computation 3.1: **a test of reliability 0.5, lengthened step by step.**

length	reliability	short of 1 by	approximation $(1-\rho)/(n\rho)$
10 times	**0.90909091**	**0.09090909**	**0.10000000**
100 times	**0.99009901**	**0.00990099**	**0.01000000**
10000 times	**0.99990001**	**0.00009999**	**0.00010000**

> **Observation 3.2**
>
> **reliability approaches 1 but does not reach it.The shortfall approaches the approximation $(1-\rho)/(n\rho)$, lying between 0.999 and 1 times it at ten thousand times the length——"measure more and it becomes accurate" is right, but only asymptotically.**

4. What Is Proportional to Length Is the Odds

Computation 4.1: **the odds of reliability $\rho/(1-\rho)$ compared with the factor of length (the original odds are 1).**

length	odds
2 times	**2.000000**
9 times	**9.000000**
100 times	**100.000000**

> **Observation 4.2**
>
> **the odds are exactly proportional to length (all three to 10^{-9}).The odds are the ratio of true-score variance to error variance, and multiplying the questions by n multiplies the first by n^2 and the second by n——this restates the result of Paper 334 that combining independent measurements adds precisions, and shares its root. Reliability seems to stick near 1 only because it is a bounded quantity; seen as odds, its growth is uniform.**

5. A Gap in What Is Measured

Computation 5.1: **the true correlation with what is wanted, Y, set at 0.6, with Y measured without error, as the test is lengthened.**

length	reliability	observed correlation
1 time	**0.500000**	**0.424264**
9 times	**0.900000**	**0.569210**
100 times	**0.990099**	**0.597022**
without limit	approaches **1**	approaches **0.600000**

Observation 5.2

this is the core.Lengthening removes only chance error; the true correlation of 0.6, that is the gap between what is measured and what is wanted, is not reduced at all by length——a hundred times the questions gives 0.597022, stopping short of 0.6. The remaining distance comes not from error but from measuring something different.

Observation 5.3

if the measure of what is wanted has error too, the ceiling drops further.With its reliability at 0.7, the observed correlation cannot exceed 0.501996 however long the test——polishing one measure leaves the error of the other.

Observation 5.4

the standard error of measurement is 0.707107 at reliability 0.5 and 0.316228 at 0.9 (taking the standard deviation of scores as 1).Nine times the length leaves one person's score swinging a little less than half as much——the swing shrinks, but its centre stays displaced from what is wanted.

6. The Separator

Observation 6.1

the separator is chance error against a gap in what is measured.Chance error vanishes on average when the same thing is measured again and again——a gap in what is measured remains however often it is measured. Length removes only the first.

Observation 6.2

the root of Paper 334, and a step beyond.It showed that combining independent measurements adds precisions——the proportionality of the odds here restates it. What this paper adds is that what is added up is precision against error only, with no power to bring what is measured closer to what is wanted.

Observation 6.3

the shape of Paper 470.There, a test ninety-nine per cent accurate could be wrong nine times in ten——here too a reliability of 0.99 says how closely repeated measurements of one thing agree, not how close that thing is to what is wanted. Stating a test's quality as one number drops what it is good for.

Observation 6.4

distinct roots, under Paper 300."Lengthening raises reliability" comes from error vanishing on average ——"correlation tops out at the true correlation" comes from the definition of what is measured. With different roots, however far the one is pushed, the other does not move.

Observation 6.5

what is not said.No exam is designed, and how long a test should be is not stated——what is shown is only the distinction between what length can buy and what it cannot.

7. Roll Call and Closing

1. **9.000000 times from 0.5 to 0.9** (Computation 2.1): **returns the target exactly.**
2. **99.000000 times to 0.99** (Computation 2.1): **the last digits cost most.**
3. **0.99990001 at ten thousand times** (Computation 3.1): **never 1.**
4. **the odds are exactly proportional to length** (Computation 4.1): **the root of Paper 334.**
5. **observed correlation from 0.424264 to 0.569210** (Computation 5.1): **lengthening raises it.**
6. **yet it tops out at 0.6** (Observation 5.2): **the core.**
7. **0.501996 if the other measure has reliability 0.7** (Observation 5.3): **the other error remains.**
8. **standard error from 0.707107 to 0.316228** (Observation 5.4): **the swing shrinks, the centre stays off.**
9. **the separator is chance error or a gap in what is measured** (Observation 6.1): **length removes only the first.**
10. **the shape of Paper 470** (Observation 6.3): **what it is good for is dropped.**

Closing: **When a test's reliability is low, the advice is "add more questions", and it is right.But what adding questions removes is only chance error——raising reliability 0.5 to 0.9 takes 9.000000 times the length and to 0.99 takes 99.000000 times, and even 10000 times gives 0.99990001, never 1. What is exactly proportional to length is not**

the reliability but its odds, which shares its root with the result of Paper 334 that combining independent measurements adds precisions.The correlation with what is wanted tops out at the true correlation——with a true correlation of 0.6 the observed correlation rises from 0.424264 to 0.569210 and 0.597022, but never exceeds 0.6 however long the test; if the measure of what is wanted has reliability 0.7, the ceiling drops to 0.501996. The standard error of measurement falls from 0.707107 to 0.316228, yet the centre of the swing stays displaced from what is wanted.The separator is chance error against a gap in what is measured——chance error vanishes on average when measurement is repeated; a gap in what is measured remains however often it is repeated. Under Paper 300 the two have distinct roots, and lengthening does not move the second.Placed among the earlier papers ——it shares the root of Paper 334 and says what lies beyond it: what is added up is only precision against error. As in Paper 470, stating a test's quality as one number drops what it is good for.To be honest——parallel questions were assumed, without fatigue or practice, and the true correlation and the other measure's reliability are representative values. And no exam is designed: what is shown is the distinction between what length can buy and what it cannot.

References

1. Brown, W. (1910). *Some Experimental Results in the Correlation of Mental Abilities*. British Journal of Psychology.
2. Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.
3. Spearman, C. (1910). *Correlation Calculated from Faulty Data*. British Journal of Psychology.
4. Yamagishi, Y. (2026). *The Three Hundredth Census: Whether Two Things Share a Root Is Decidable* [Paper 300]. Zenodo. 10.5281/zenodo.22171914
5. Yamagishi, Y. (2026). *What Is Being Added Is Not the Values but the Precisions* [Paper 334]. Zenodo. 10.5281/zenodo.22681804
6. Yamagishi, Y. (2026). *A Test That Is Ninety-Nine Per Cent Accurate Can Be Wrong Nine Times in Ten* [Paper 470]. Zenodo. 10.5281/zenodo.22733434

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)